Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2015, Article ID 343869, 15 pages
http://dx.doi.org/10.1155/2015/343869

Research Article

Anomaly Detection via Midlevel Visual Attributes

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

[1]*Key Laboratory of Machine Perception, Peking University, Beijing 100084, China*
[2]*CRSC Communication & Information Corporation, Beijing 100070, China*

Correspondence should be addressed to Tan Xiao; pkuxiaotan@pku.edu.cn

Received 23 August 2014; Revised 18 November 2014; Accepted 24 November 2014

Academic Editor: Yi Jin

Automatically discovering anomalous events and objects from surveillance videos plays an important role in real-world application and has attracted considerable attention in computer vision community. However it is still a challenging issue. In this paper, a novel approach for automatic anomaly detection is proposed. Our approach is highly efficient; thus it can perform real-time detection. Furthermore, it can also handle multiscale detection and can cope with spatial and temporal anomalies. Specifically, local features capturing both appearance and motion characteristics of videos are extracted from spatiotemporal video volume (STV). To bridge the large semantic gap between low-level visual feature and high-level event, we use the middle-level visual attributes as the intermediary. And these three-level framework is modeled as an extreme learning machine (ELM). We propose to use the spatiotemporal pyramid (STP) to capture the spatial and temporal continuity of an anomalous even, enabling our approach to cope with multiscale and complicated events. Furthermore, we propose a method to efficiently update the ELM; thus our approach is self-adaptive to background change which often occurs in real-world application. Experiments on several datasets are carried out and the superior performance of our approach compared to the state-of-the-art approaches verifies its effectiveness.

1. Introduction

1.1. Motivation. Surveillance systems have been widely used in the city, and detecting anomalous events from the system plays an important role in real world. However, current systems are actually quite burdensome for human operators because they are required to watch a large quantity of screens (usually up to 50 [1]) which show the content captured by several different cameras. Detecting unusual individuals and events [2], a.k.a., anomalies, is one of the most important and main tasks of human operators. Thus the performance of anomaly detection is highly dependent on the human operators. However, the quantity of cameras is growing explosively in the city requiring more and more human operators, and the task is becoming more difficult and tiring for human operators such that the performance of operators can degrade significantly [3]. Fortunately, with the development of video analytics techniques, automatic anomaly detection approaches, which can analyze video streams automatically to warn, possibly in real time, the human operators that an anomalous event is currently taking place, have attracted considerable attention in recent years.

Specifically, anomaly detection is defined as discovering events with a low probability of occurrence from surveillance videos in computer vision community. Several approaches have been proposed in recent years and generally they can be summarized in the following three categories based on how their models are constructed, that is, supervised approaches [4–10], semisupervised approaches [11], and unsupervised approaches [12–21]. In real-world scenario, anomalies are usually quite rare, as its definition. And they show significant difference between each other and they also have unpredictable variations. Thus unsupervised approaches are more favored in most recent years. Moreover, since anomalous events are indeed difficult, almost impossible to be defined in advance, unsupervised approaches are actually practical in real-world applications.

Furthermore, unsupervised approaches can also be divided into several subcategories according to the techniques they use. Trajectories based approaches [17, 22–24]

focus on the spatial location of objects or persons and their motions are tracked. But these approaches have a significant limitation that they can only capture the abnormal track because they only consider the spatial deviations; thus an abnormal target with abnormal appearance and motion but following a normal track cannot be detected. In addition, because precise segmentation of a target is almost impossible in a crowd scene, these approaches cannot be applied to such scenario. To model typical motion of objects, optical flow has been widely utilized also [15, 16]. But as mentioned in [22], these approaches have very unstable performance in crowded scenes. And they can only detect anomalous motion while ignoring anomalous appearance because they just focus on the motion of objects and persons. Recently, densely sampled local spatiotemporal descriptor which represents both motion and appearance characteristics are utilized in [20, 21], and they can also possess some degree of robustness to unimportant variations in surveillance video. They construct models to capture the relationship between low-level visual features and high-level semantic event. Though promising results are achieved, they ignore the significant semantic gap between low-level visual features and high-level events; thus their performance is still unsatisfactory for real-world application.

Summarized from previous works, an effective and applicable anomaly detection approach should satisfy the following properties. (1) It should be definitely unsupervised because defining all anomalous events in advance is almost impossible and too burdensome for human operators to do so. (2) Both spatial and temporal anomalies, that is, both anomalies of motion and appearance, can be detected by the approach. (3) It can detect multiscale anomalies because it is also hard to know a priori the range of an anomaly, for example, its size, speed, and duration. (4) It can be online updated to self-adapt to scene change in both motion and appearance. Actually, the appearance of background is always changing in surveillance videos because of lighting condition, weather, and so forth. And the change of motion pattern should not be ignored too. (5) Last but the most important, it is able to detect the anomalies from the surveillance videos effectively and efficiently.

1.2. Contribution. To address this challenging problem, in this paper we propose a novel automatic anomaly detection approach with extreme learning machine (ELM) [25] based visual attribute and spatiotemporal pyramid (STP). The former one focuses on the relationship between low-level visual features and high-level event and the latter can capture the spatial and temporal continuity of an event. We propose to combine them because both parts are important for anomaly detection.

Specifically, spatiotemporal video volumes (STV) with pixel-by-pixel analysis which are densely sampled from surveillance video lay the foundation of our approach. Then spatiotemporal feature descriptor, which can capture both motion and appearance characteristics, is extracted for each STV, and each STV is further represented by bag-of-video-words HOG feature. Then to bridge the semantic gap between low-level visual features and high-level event, we propose to use visual attributes as the intermediary, which is motivated by the extensive research on attribute learning for visual analysis recently [26–28]. We propose to model this three-level (feature-attribute-event) framework by extreme learning machine (ELM). The output of the ELM can be utilized to tell whether the STV belongs to an anomaly. And since the ELM can be constructed and updated with extremely high efficiency, the model can be updated continuously with coming surveillance videos; thus no offline or supervised pretraining is required for our model. So our approach can detect anomaly which even has not been observed before. And the efficient update procedure also enables our approach to cope with the scene change in both motion and appearance. Furthermore, to detect multiscale and complicated anomalies, we use spatiotemporal pyramid (STP), which is the temporal extension of spatial pyramid [29]. Thus, with different scales, multiscale event can be effectively detected by STP. Moreover, since an event is always related to several STVs which may have different location or time or both, complicated events can be detected by discovering the relationship. STP can achieve this task, enabling our approach to detect complicated events. Thus the proposed approach can effectively and accurately detect multiscale and complicated anomalies, both in motion and appearance. And the model can be quite efficiently constructed and updated in an unsupervised way; thus it can detect anomalies which have not been observed before. Moreover, our approach is self-adaptive to background change in both motion and appearance.

We show the framework overview of the proposed approach in Figure 1. The input is a video stream. Then the stream is densely sampled; that is, it is sampled pixel by pixel. Around each pixel, a 3D volume is constructed, that is, STV. This volume is segmented with no overlap into 8 smaller STVs. Thus the upper and coarser level of STP is formed by the larger STV which can capture the overall information of an event. And the smaller STVs form the lower but finer level of STP, which can capture the details of an event. Actually, we can continue segmenting any small STV into another 8 smaller STVs. But this leads to much more computational complexity and we find that a two-level STP can achieve satisfactory performance for anomaly detection. Then, for each STV, HOG features which can capture both motion and appearance characteristics of STV are extracted. The HOG of upper level STV can be efficiently constructed from its lower level STVs; thus our multilevel STP actually does not require too much extra computation for feature extraction, which is an essential property for real-time detection. Instead of learning a model that directly connect low-level visual features to high-level event which may suffer from the large sematic gap between them, we propose to use visual attribute [26] as the middle level to bridge the semantic gap. And we utilize extreme learning machine to model this three-level framework; that is, low-level feature is the input, visual attribute is the hidden unit, and the output of ELM can be regarded as the high-level event. And ELM can be efficiently trained and updated which is also important. Finally, the anomaly judgement is given based on the results of both levels.

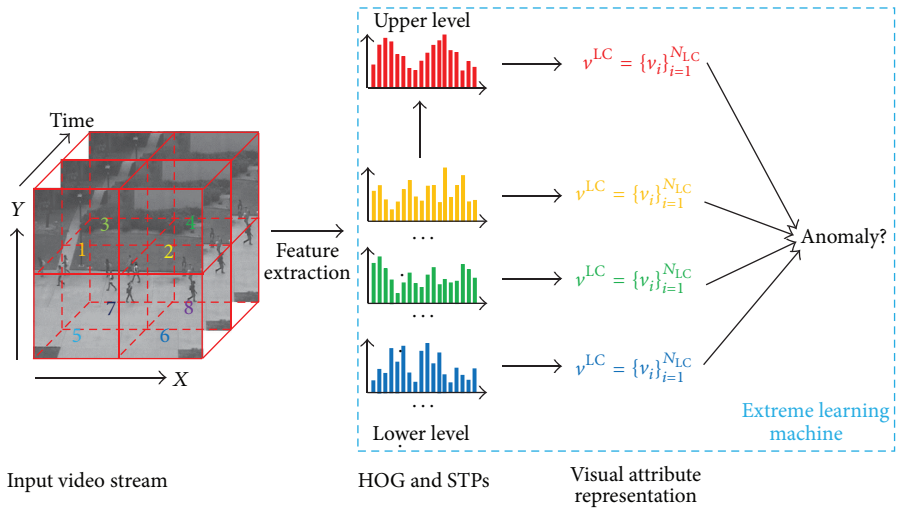

FIGURE 1: Overview of our approach. This is an example of two-level spatiotemporal pyramid. The input is a video stream. Then a 3D volume around a pixel is constructed represented by the outer red cube. Then it is segmented into 8 ($2 \times 2 \times 2$) smaller cubes denoted by different numbers in 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. We use visual attribute representation to bridge the semantic gap between low-level feature and high-level event. The three-level (feature-attribute-event) framework can be modeled by extreme learning machine. Finally the anomaly detection is completed by combining the outputs of the machine.

In a summary, we make the following contributions to this paper.

(i) A novel approach for automatic anomaly detection is proposed. It is based on densely sampled STVs. Visual attribute is utilized to bridge the semantic gap and we use ELM to model this three-level framework. ELM can also effectively and efficiently tell whether a STV belongs to an anomalous event.

(ii) We propose to use spatiotemporal pyramid (STP) to capture the spatial and temporal continuity of an anomalous event. And our approach can perform multiscale detection with the presence of STP.

(iii) The use of ELM has another important benefit; that is, it can be efficiently updated. Thus our approach can perform efficient online update and thus can adaptively learn the event patterns in the scene and cope with the scene change of both motion and appearance.

(iv) Extensive experiments on several public datasets are carried out to evaluate the effectiveness of our approach for anomaly detection. And the results show that our approach can achieve satisfactory performance and significantly outperform several state-of-the-art approaches.

The rest of this paper is organized as follows. In Section 2, we will briefly review some works related to our approach. The details for detection via ELM-based visual attribute and online updating method are described in Section 3. We will introduce the spatiotemporal pyramid in Section 4. The experiments and results are presented in Section 5 and we draw conclusions in Section 6.

2. Related Work

As has been mentioned in Section 1, one of the most widely used techniques in previous works is trajectory analysis. But generally, precise tracking methods [30, 31] are always required by this kind of methods. But unfortunately, tracking objects is quite time-consuming and computationally expensive, especially in crowded scenes where a large number of objects and persons are moving such that precise segmentation of targets is nearly impossible, which is the foundation of tracking analysis. And some previous works also utilize optical flow [15, 16], but their performance in crowded scenes is quite unreliable [20].

Recently, the focus of anomaly detection has turned from object detection or tracking to local spatiotemporal features. Several approaches have been proposed and received increasing attention [32, 33]. Typically, these approaches focus on pixel level. They propose to describe the local characteristic of video by low-level visual features, like color, texture, and motion. Then they can construct pixel-level background model and behavior template based on the local features [34–37]. In addition, approaches utilizing spatiotemporal video volumes in the context of bag-of-video-words have achieved promising results [13, 20, 38]. For example, probabilistic models such as latent Dirichlet allocation (LDA) [39] can be straightforwardly applied to video analysis if we ignore the spatial and temporal relationship of local features [40, 41]. But intuitively, the spatial and temporal relationship between STVs are quite essential for scene understanding and event detection [42]. Noticing this point, the efforts to incorporate either spatial or temporal composition of STVs into the conventional probabilistic model have been made in some works. But they are not able to handle online and real-time detection because they are highly time-consuming

and computationally expensive [40]. In addition, several approaches [32, 35, 43, 44] try to construct models based on the spatiotemporal behavior and analyse the spatiotemporal pattern of each pixel as a function of time to detect low-level local anomalous events. However, they ignore the relationship between each pixel in space and time because they just process each pixel independently such as in [43], which may lead to too local detection.

A multiscale and nonparametric approach is proposed in [20] to perform real-time anomaly detection and localization. Dense and local spatiotemporal features which can capture both motion and appearance characteristics of objects are extracted at each individual pixel. And to take advantage of the spatial and temporal relationship between pixels, "overlapping" features are utilized in their approach. As reported in their paper, they can achieve promising results, but their approach indeed faces the challenge of efficiency when performing accurate multiscale anomaly detection. Actually, our spatiotemporal pyramid is partially motivated by their overlapping features because we both consider the spatial and temporal relationship between features. But compared to their overlapping features, our STP can be constructed with much more efficiency and much better performance can be achieved. Furthermore, our STP can cope with multiscale detection naturally while their approach actually treats different scales independently.

Moreover, approaches mentioned above all construct models between low-level visual features and high-level events straightforwardly while ignoring the semantic gap between them which is quite important for event detection. Thus they cannot achieve satisfactory results. This problem motivates us to use visual attribute as intermediary to bridge the semantic gap for more accurate detection.

3. Detection via ELM-Based Visual Attribute

3.1. Spatiotemporal Local Features. Tracking objects or persons in videos is quite time consuming and its performance is unstable under crowded scene. Instead, we utilize the local features which can capture the local spatial and temporal characteristics of video. By analyzing the pattern of spatiotemporal local features, we can detect and localize the anomalies.

First of all, we need to extract meaningful local features to capture the motion and appearance characteristics of densely sampled STVs at each pixel. Considering a pixel (x, y, t), we can construct a STV $v \in \mathbb{R}^{n_x \times n_y \times n_t}$ with the size $n_x \times n_y \times n_t$ centered at (x, y, t), where $n_x \times n_y$ denotes the size of spatial window and n_t is the depth of STV in time. Typically, $5 \times 5 \times 5$ or $10 \times 10 \times 10$ can be proper size of a STV. Then we calculate the histogram of the spatiotemporal gradient (HOG) of the video in polar coordinates to describe the STV, following the works in [20, 21, 45]. Now we can 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). Specifically, they are computed using the finite difference approximations as follows:

$$\begin{aligned} G_x(x, y, t) &= L_{\sigma_d}(x + 1, y, t) - L_{\sigma_d}(x - 1, y, t), \\ G_y(x, y, t) &= L_{\sigma_d}(x, y + 1, t) - L_{\sigma_d}(x, y - 1, t), \\ G_t(x, y, t) &= L_{\sigma_d}(x, y, t + 1) - L_{\sigma_d}(x, y, t - 1), \end{aligned} \tag{1}$$

where L_{σ_d} is obtained by filtering the signal with a Gaussian kernel of bandwidth σ_d to suppress the noise. In our experiment, we find out that setting $\sigma_d = 1.1$ leads to satisfactory performance.

In real-world applications, surveillance videos are always affected by noise and the effect of local texture and contrast also have significant influence on the video analysis. Thus, in order to alleviate the influence of noise in videos and texture and contrast, we first normalize the spatial gradient as follows:

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

where $G_s(x, y, t)$ is the normalized spatial gradient and ϵ is a small constant to avoid numeric instabilities (denominator is equal to zero by chance). Typically we can set $\epsilon = 0.01$. Based on the normalized spatial gradient, we can further construct 3D normalized gradient represented in polar coordinates as follows:

$$\begin{aligned} M_{3D}(x, y, t) &= \sqrt{G_s^2(x, y, t) + G_t^2(x, y, t)}, \\ \theta(x, y, t) &= \tan^{-1}\left(\frac{G_y(x, y, t)}{G_x(x, y, t)}\right), \\ \phi(x, y, t) &= \tan^{-1}\left(\frac{G_t(x, y, t)}{G_s(x, y, t)}\right), \end{aligned} \tag{3}$$

where $M_{3D}(x, y, t)$ is the magnitude of 3D normalized gradient and $\phi(x, y, t) \in [-\pi/2, \pi/2]$ and $\theta(x, y, t) \in [-\pi, \pi]$ are the orientations of the gradient, respectively. Then we can construct the histogram of oriented gradients (HOG features) for a given STV v in the following way. First, for each pixel in the give STV v, we can extract 3D normalized gradient features for them. Then the feature for each pixel is quantized into $n_\phi + n_\theta$ bins based on their gradient orientations. Typically we can set $n_\phi = 8$ and $n_\theta = 16$. So the HOG features for STV v, denoted by h, has 24 dimensions under this setting. If we look back to the feature extraction and construction procedure, we can observe that the local characteristics of both motion and appearance in the video can be captured by the HOG features. Consequently, both anomalous actions and objects can be detected based on the HOG features. Moreover, because of the normalization step at first, it shows robustness to data noise and unimportant variations in the video such as texture and contrast. Actually, this HOG feature can be used as the input of the extreme learning machine.

Besides the effectiveness of low-level visual features, we also care about the efficiency of feature extraction and construction. Though it seems that the feature construction is

quite computationally expensive, actually we can notice that it is quite efficient because the computation can be always reused as discussed below.

For example, the 3D normalized gradient at one pixel will be used by $n_x \times n_y \times n_t$ times because it is used to construct HOG features for any STV containing it. If the 3D normalized gradient is repeatedly computed for all STVs, it is indeed burdensome. But we can precompute it in advance and we just need to reuse the obtained result for each STV. Also, the histogram of pixel (x, y, t), denoted by $h(x, y, t)$, can also be precomputed by quantization a priori. Then the histogram of a STV v around pixel (x, y, t) can be computed by simply summing up all the histograms in this STV as

$$h_v(x, y, t) = \sum_{(x', y', t') \in v} h(x', y', t'). \tag{4}$$

It is obvious that just summing up all the histograms can be quite efficient. On the other hand, if we compute the HOG feature for each STV independently, it can be highly computational too. But we can observe that we can use its neighbor's HOG feature to construct the HOG feature for a given STV as follows; thus computations can be saved markedly:

$$\begin{aligned} h_{v_2} = h_{v_1} &- \sum_{(x', y', t') \in v_1 \setminus v_2} h(x', y', t') \\ &+ \sum_{(x', y', t') \in v_2 \setminus v_1} h(x', y', t'), \end{aligned} \tag{5}$$

where "\" is the set minus operation and the STVs around (x, y, t) and $(x+1, y, t)$ are denoted by v_1 and v_2, respectively. It is clear that $v_1 \setminus v_2$ is much smaller than v_1. Moreover, 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, that is, the red cube, can be computed as follows:

$$\begin{aligned} h_{v_{\text{up}}} &= \sum_{(x', y', t') \in v_{\text{up}}} h(x', y', t') \\ &= \sum_{i=1}^{8} \sum_{(x', y', t') \in v_i} h(x', y', t') = \sum_{i=1}^{8} h_{v_i}. \end{aligned} \tag{6}$$

Consequently, because of the computational tricks above, the extraction and construction of the spatiotemporal feature can be highly efficient, which is one of the most important requirements for real-time detection.

3.2. Constructing Normal Events. As mentioned in Section 1, our approach is unsupervised. Thus we need to define normal and anomalous events automatically. As the definition of anomaly, it is quite rare compared to normal events. Thus, the number of STVs associated with anomalies is much fewer than the number of STVs belonging to normal events in video. Generally, STVs belonging to normal events may form clusters in the feature space while STVs of anomalous events are outliers in the space. Thus we can construct normal events by using this property which is illustrated in Figure 2. Consequently, we can regard the clusters as the normal events and the judgement of anomaly can be given by analyzing how close a STV is related to a cluster. Intuitively, we can construct these clusters by some clustering methods such us k-means. However, there are some parameters for clustering algorithm; for example, we need to specify k for k-means clustering. And we find out that our approach is a little sensitive to this parameter.

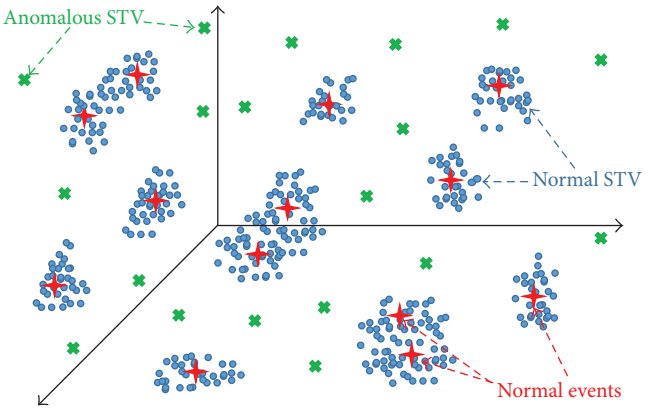

Figure 2: Constructing normal events. Generally, STVs belonging to normal events may form clusters in the feature space while STVs of anomalous events are outliers in the space. We can use this property to construct normal events.

Instead, we propose to construct normal events automatically from video data. Given a set of spatiotemporal features $\mathbf{H} = [h_1, \ldots, h_n] \in \mathbb{R}^{d \times n}$, where $d = 24$ is the dimension of feature and n is the size of feature set. Actually we do not need a training set because the initial feature set $\mathbf{H}$ can be constructed by using the first few seconds of the video. In addition, to guarantee that our selection algorithm is computationally feasible, we can also randomly select some features to reduce n. In real-world scenario, n can be tuned from 10,000 to 20,000 according to the resolution of videos. Then we need to select some features from $\mathbf{H}$ as the representatives of clusters, that is, the normal events. In our method, the number of clusters is determined automatically by the algorithm, which is self-adaptive to the test data. Following the idea in [19], we would like to select an optimal subset of $\mathbf{H}$ as the clusters, such that we can well reconstruct the rest of features from them. This criterion can be formulated as follows:

$$\min_{\mathbf{S}} = \frac{1}{2} \|\mathbf{H} - \mathbf{HS}\|_F^2 + \lambda \|\mathbf{S}\|_{2,1}, \tag{7}$$

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 matrix $\mathbf{S}$, $\|\mathbf{S}\|_{2,1} = \sum_{i=1}^{n} \|\mathbf{S}_i\|_2$ is the $L_{2,1}$-norm of matrix, and λ is the model parameter. Finally, the selection can be done by selecting index i which satisfies $\|\mathbf{S}_i\| > 0$. Consequently we can obtain clusters representing the normal events.

To solve this problem, we follow the method proposed in [46]. Consider an objective function $f_0(x) = f(x) + g(x)$ where $f(x)$ is convex and smooth and $g(x)$ is convex but nonsmooth. The key step is to construct $p_{Z,L}(x) = f(Z) + \langle \nabla f(Z), x - Z\rangle + (L/2)\|x - Z\|_F^2 + g(Z)$ to approximate $f_0(x)$ at point Z. Obviously, we can define $f(\mathbf{S}) = (1/2)\|\mathbf{H} - \mathbf{HS}\|_F^2$ and $g(\mathbf{S}) = \lambda\|\mathbf{S}\|_{2,1}$. So we can construct $p_{\mathbf{Z},L}(\mathbf{S})$ as

$$p_{\mathbf{Z},L}(\mathbf{S}) = f(\mathbf{Z}) + \left\langle \nabla f(\mathbf{Z}), \mathbf{S} - \mathbf{Z}\right\rangle + \frac{L}{2}\|\mathbf{S} - \mathbf{Z}\|_F^2 + g(\mathbf{Z}). \tag{8}$$

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

$$\mathbf{N}_{i.} = \begin{cases} 0, & \|\mathbf{M}_{i.}\| \le \tau \\ \left(1 - \dfrac{\tau}{\|\mathbf{M}_{i.}\|}\right)\mathbf{M}_{i.}, & \text{otherwise.} \end{cases} \tag{9}$$

Theoretically, $\mathbf{S}$ is given by solving the following optimization problem:

$$\mathbf{S} = \arg\min_{\mathbf{S}} p_{\mathbf{Z},L}(\mathbf{S}). \tag{10}$$

And this optimization problem can be equivalently written as follows:

$$\begin{aligned} &\min_{\mathbf{S}} f(\mathbf{Z}) + \left\langle \nabla f(\mathbf{Z}), \mathbf{S} - \mathbf{Z}\right\rangle + \frac{L}{2}\|\mathbf{S} - \mathbf{Z}\|_F^2 + \lambda\|\mathbf{S}\|_{2,1} \\ &\iff \min_{\mathbf{S}} \frac{L}{2}\left\|(\mathbf{S} - \mathbf{Z}) + \frac{1}{L}\nabla f(\mathbf{Z})\right\|_F^2 + \lambda\|\mathbf{S}\|_{2,1} \\ &\iff \min_{\mathbf{S}} \frac{L}{2}\left\|\mathbf{S} - \left(\mathbf{Z} - \frac{1}{L}\nabla f(\mathbf{Z})\right)\right\|_F^2 + \lambda\|\mathbf{S}\|_{2,1} \\ &\iff \min_{\mathbf{S}} \frac{L}{2}\left\|\mathbf{S} - \left(\mathbf{Z} - \frac{1}{L}\nabla f(\mathbf{Z})\right)\right\|_F^2 + \lambda\sum_{i=1}^{n}\|\mathbf{S}_{i.}\|_2. \end{aligned} \tag{11}$$

And since the L_2 norm is self-dual, the problem can be further rewritten as follows by introducing a dual variable $\mathbf{Y} \in \mathbb{R}^{n\times n}$:

$$\begin{aligned} &\min_{\mathbf{S}} \frac{L}{2}\left\|\mathbf{S} - \left(\mathbf{Z} - \frac{1}{L}\nabla f(\mathbf{Z})\right)\right\|_F^2 + \lambda\sum_{i=1}^{n}\max_{\|\mathbf{Y}_{i.}\|\le 1}\langle \mathbf{Y}_{i.}, \mathbf{S}_{i.}\rangle \\ &\iff \max_{\|\mathbf{Y}_{i.}\|\le 1}\min_{\mathbf{S}} \frac{L}{2}\left\|\mathbf{S} - \left(\mathbf{Z} - \frac{1}{L}\nabla f(\mathbf{Z})\right)\right\|_F^2 + \lambda\sum_{i=1}^{n}\langle \mathbf{Y}, \mathbf{S}\rangle \\ &\iff \max_{\|\mathbf{Y}_{i.}\|\le 1}\min_{\mathbf{S}} \frac{L}{2}\left\|\mathbf{S} - \left(\mathbf{Z} - \frac{1}{L}\nabla f(\mathbf{Z}) - \frac{\lambda}{L}\mathbf{Y}\right)\right\|_F^2 \\ &\qquad - \frac{1}{2}\left\|\mathbf{Z} - \frac{1}{L}\nabla f(\mathbf{Z}) - \frac{\lambda}{L}\mathbf{Y}\right\|_F^2. \end{aligned} \tag{12}$$

The second equation above is obtained by swapping "min" and "max." Because this function is convex with respect to $\mathbf{S}$ and concave with respect to $\mathbf{Y}$, this swapping will not change the problem by Von Neumann minimax theorem. Further, denote $\mathbf{S} = \mathbf{Z} - (1/L)\nabla f(\mathbf{Z}) - (\lambda/L)\mathbf{Y}$. From the last equation above we can obtain an equivalent problem as follows:

$$\max_{\|\mathbf{Y}_{i.}\|\le 1} -\frac{1}{2}\left\|\mathbf{Z} - \frac{1}{L}\nabla f(\mathbf{Z}) - \frac{\lambda}{L}\mathbf{Y}\right\|_F^2. \tag{13}$$

Analogous to the substitution above, let $\mathbf{Y} = -(L/\lambda)(\mathbf{S} - \mathbf{Z} + (1/L)\nabla f(\mathbf{Z}))$; we can change the problem above into a problem in terms of the original variable $\mathbf{S}$ as

$$\begin{aligned} &\min_{\|(L/\lambda)(\mathbf{S}-\mathbf{Z}+(1/L)\nabla f(\mathbf{Z}))_{i.}\|_F^2\le 1} \|\mathbf{S}\|_F^2 \\ &\iff \sum_{i=1}^{n}\min_{\|\mathbf{S}_{i.}-(\mathbf{Z}-(1/L)\nabla f(\mathbf{Z}))_{i.}\|_2\le\lambda/L}\|\mathbf{S}_{i.}\|_2^2. \end{aligned} \tag{14}$$

Therefore, the optimal solution to the first problem above is equivalent to the last problem above. Actually, we can optimize each row in $\mathbf{S}$ independently in the last problem. Considering each row of **S,** respectively, we can get the closed form as

$$\mathbf{S} = \arg\min_{\mathbf{S}} p_{\mathbf{Z},L}(\mathbf{S}) = D_{\lambda/L}\left(\mathbf{Z} - \frac{1}{L}\nabla f(\mathbf{Z})\right). \tag{15}$$

We summarize the whole algorithm in Algorithm 1.

3.3. Detection via ELM-Based Visual Attributes. In the above two subsections, we introduce how to extract low-level visual features which can simultaneously capture both motion and appearance characteristics of videos and how to construct a set of normal events in an unsupervised way. Intuitively, one simple way to tell whether a STV represented by low-level visual feature belongs to a anomaly is to consider the relationship between the low-level feature and all high-level normal events. In fact, if it is closely related to one event, it has very high probability to be normal. On the contrary, if it does not have strong relation to any one normal events, it usually belongs to an anomaly. Thus we just need to define a function or model to measure the relationship between low-level visual features and high-level events. However, directly constructing model between low-level visual features and high-level events may suffer from the large semantic gap between them. Recent research on visual attribute [26–28] has pointed out this problem and proposed to use visual attribute as the intermediary to bridge the semantic gap.

The basic idea is shown in Figure 3. Instead of constructing model between low-level visual features and high-level events directly, now we need construct two models, that is, one between low-level visual features and middle-level visual attributes and one between middle-level visual attributes and high-level events. Actually, directly connecting low-level visual features and high-level events in only one step is quite difficult. Thus we use two steps such that either one can be more feasible.

Theoretically, we can apply any models to both steps. And we find that linear model can always lead to satisfactory performance while guaranteeing the efficiency for both training and detection given a set of STVs and the corresponding low-level visual features $\mathbf{H} = [h_1, \ldots, h_n] \in \mathbb{R}^{d\times n}$, and the

Input: $\mathbf{H}$, $\lambda = 1$, $\mathbf{S}_0$, K, c
Output: $\mathbf{S}$;
(1) Initialize $\mathbf{Z}_0 = \mathbf{S}_0$, $a_0 = 1$
(2) **for** $k = 0, 1, 2, \ldots, K$ **do**
(3) $\mathbf{S}_{k+1} = \arg\min_{\mathbf{S}} : p_{\mathbf{Z}_k,L}(\mathbf{S}) = D_{\lambda/L}\left(\mathbf{Z}_k - \frac{1}{L}\nabla f(\mathbf{Z}_k)\right)$
(4) **while** $f_0(\mathbf{S}_{k+1}) > p_{\mathbf{Z}_k,L}(\mathbf{S}_{k+1})$ **do**
(5) $L = \frac{L}{c}$
(6) $\mathbf{S}_{k+1} = \arg\min_{\mathbf{S}} : p_{\mathbf{Z}_k,L}(\mathbf{S}) = D_{\lambda/L}\left(\mathbf{Z}_k - \frac{1}{L}\nabla f(\mathbf{Z}_k)\right)$
(7) **end while**
(8) $a_{k+1} = \frac{\left(1 + \sqrt{1 + 4a_k^2}\right)}{2}$
(9) $\mathbf{Z}_{k+1} = \left(\frac{a_{k+1} + a_k - 1}{a_{k+1}}\right)\mathbf{S}_{k+1} - \left(\frac{a_k - 1}{a_{k+1}}\right)\mathbf{S}_k$
(10) **end for**

Algorithm 1: Normal events construction.

corresponding event labels $\mathbf{E} = [e_1, \ldots, e_n] \in \{-1, 1\}^{m \times n}$, where $e_{ij} = 1$ if the jth STVs belongs to the ith event and $e_{ij} = -1$ otherwise. Previous approaches directly construct models between $\mathbf{H}$ and $\mathbf{E}$ which may suffer from the semantic gap between them. Here we propose to utilize a middle-level $\mathbf{A} = [a_1, \ldots, a_n] \in \mathbb{R}^{k \times n}$ to bridge the semantic gap, where k is the number of visual attributes which we set as $k = 256$ in this paper. Actually, a_j can be regarded as the visual attribute for the ith STV. Thus we can construct two linear models as follows:

$$a_{ij} = g\left(w_i \cdot h_j + b_i\right), \quad i = 1, \ldots, k$$
$$e_{ij} = \sum_{t=1}^{k} \beta_{it} \cdot a_{tj}, \quad i = 1, \ldots, m, \tag{16}$$

where a_{ij} is the ith attribute for the jth STV and e_{ij} is the relation degree between the ith event and the jth STV which is used to show that the STV belongs to an anomaly or not. Now we just need to specify two model parameters, that is, $\mathbf{W} = [w_1, \ldots, w_k] \in \mathbb{R}^{d \times k}$ and $\mathbf{B} = [b_1, \ldots, b_k]^T \in \mathbb{R}^k$ for the model between low-level visual feature and middle-level visual attribute and $\beta = [\beta_1, \ldots, \beta_k] \in \mathbb{R}^{m \times k}$ for the model between middle-level visual attribute and high-level event. And $g(\cdot)$ is the activation function which is infinitely differentiable, such as sigmoidal function, radial basis, cosine, and exponential. In this paper we use sigmoidal function as the activation function.

Now we need to minimize the following objective function to determine model parameters:

$$O = \sum_{j=1}^{n} \left\| \sum_{i=1}^{k} \beta_i \cdot g\left(w_i \cdot h_j + b_i\right) - e_j \right\|_F^2, \tag{17}$$

where n is the number of training data. As mentioned before, our approach actually does not need prepared training data. We can use the normal event set constructed above as the training data. Suppose finally we select n events with

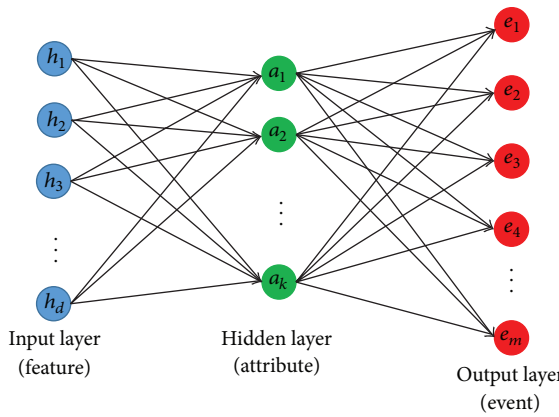

Figure 3: ELM-based visual attribute. There is large semantic gap between low-level visual feature and high-level event. To bridge the gap, we can use visual attribute as intermediary. And this three-level (feature-attribute-event) framework can be formulated as an extreme learning machine.

corresponding feature representation. Then the features can be used as $\mathbf{H}$, and each event has its own representation e_j, where $e_{jj} = 1$ and $e_{ij} = -1$ for $i \neq j$. Thus we can construct $\mathbf{H}$ and $\mathbf{E}$ as the training set in an unsupervised way while requiring no extra effort.

In fact, learning two models simultaneously is quite difficult and inefficient because we need to adjust parameters in two models iteratively, such that it cannot be applied to online and real-time applications. But fortunately, this three-level (feature-attribute-event) framework can be formulated as an extreme learning machine (ELM) [25, 47–49] which is a variant of artificial neural network (ANN), where feature is the input layer, attribute is the hidden layer, and event is the output layer. In the theory of ELM, the parameters between input layer and hidden layer can be totally random; that is, we actually do not need to learn these parameters from the training data. Thus we just need to compute the parameters between the hidden layer and the output layer,

Figure 4: Experiments on Belleview dataset.

which is extremely fast compared to conventional ANN which is solved by backpropagation.

Now we can randomize the parameters $\mathbf{W}$ and $\mathbf{B}$ for the model between low-level visual feature and middle-level visual attribute. Because we will not change $\mathbf{W}$ and $\mathbf{B}$, we can compute the visual attributes for the training data as follows:

$$\mathbf{A} = g\left(\mathbf{W}^T\mathbf{H} + \mathbf{B}\mathbf{1}_k^T\right), \tag{18}$$

where $\mathbf{1}_k = [1, \ldots, 1]^T$. Then we just need to compute β by minimizing the following objective function:

$$O = \left\|\beta\mathbf{A} - \mathbf{E}\right\|_F^2 \tag{19}$$

and we can get the solution for β as

$$\widehat{\beta} = \mathbf{E}\mathbf{A}^{\dagger}, \tag{20}$$

where $\mathbf{A}^{\dagger}$ is the Moore-Penros generalized inverse of matrix $\mathbf{A}$ [50, 51]. Then, given any STV represented by low-level visual features $h^* \in \mathbb{R}^d$, we can compute its relationship with any event as $e^* \in \mathbb{R}^m$ via the middle-level visual attribute by the ELM as follows:

$$e^* = \widehat{\beta}g\left(\mathbf{W}^T h + \mathbf{B}\right). \tag{21}$$

Now we need to discuss how to tell whether this STV belongs to an anomaly based on e^*. Actually, as discussed above, a normal STV is always close to a cluster while an abnormal STV is always outlier. And an normal event is represented by a cluster. Furthermore, e_i^* denotes the relationship between this STV with the ith event (i.e., cluster). Thus if it belongs to an normal event, it may have strong relationship with an event, leading to $e_i^* \approx 1$ for some j implying strong relationship, while no elements in e^* may have large value because it is an outlier and has quite weak relationship with all events. Consequently, we can define the degree of anomaly as

$$d_{\text{anomaly}} = 1 - \max_i\left(e_i^*\right). \tag{22}$$

Formally we can select an anomaly threshold δ such that a STV that satisfies $d_{\text{anomaly}} > \delta$ is judged as abnormal while that satisfies $d_{\text{anomaly}} \leq \delta$ is judged as normal. Now we need to determine the value of δ. We can define the anomaly probability p_a, which is empirically selected from $10^{-2}, 10^{-3}, 10^{-4}$, and 10^{-5}, depending on the user's need. High true positive rate and high false positive rate are achieved with a large p_a while a small p_a will lower both. Then we can compute d_{anomaly} 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_{anomaly} 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 in detail in Section 4.

So far, we complete the introduction to our anomaly detection algorithm based on visual attribute and extreme learning machine.

3.4. An Alternative. We construct our model based on ELM to model this three-level frame work as above. And we propose to randomize $\mathbf{W}$ and choose an infinitely differentiable activation function. The infinitely differentiable activation function always leads to continuous output, that is, real-value attribute. However, recent research demonstrates that binary attributes can lead to better performance. Thus we want the activation function to output binary value, such that we can use the sign function as the activation function defined as follows:

$$g(x) = \operatorname{sign}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ -1, & \text{otherwise.} \end{cases} \tag{23}$$

In our experiment, we find out that this activation function shows satisfactory performance as previous works [28]. However, as the sign function is not infinitely differentiable, we cannot simply randomize $\mathbf{W}$ as in last section. In this paper, we propose an alternative for learning $\mathbf{W}$ which is also quite effective and efficient, such that we can use sign function as the activation function.

First we can construct the training set with features $\mathbf{H} = [h_1, \ldots, h_n]$ and the corresponding event representations $\mathbf{E} = [e_1, \ldots, e_n]$, where $e_{jj} = 1$ and $e_{ij} = -1$ for $i \neq j$, as mentioned in the last section. Then we need to construct the binary attributes $\mathbf{A} = [a_1, \ldots, a_n] \in \{-1, 1\}^{k \times n}$ and the model parameters $\mathbf{W}$.

To construct effective visual attributes, we think two important principles should be followed: (1) the learned attributes should be predictable; that is, we can generate the attribute representation correctly from low-level visual features and (2) they should be discriminative; that is, different events should have different attribute representations.

Input: **H**, $C = \gamma = 1, k$
Output: **A**, **W**;
(1) Initialize **W** by randomization
(2) Initialize **A**: $a_j^i = \text{sign}(w_i' \cdot h_j), \forall j = 1, \ldots, n, i = 1, \ldots, k$
(3) **repeat**
(4) Optimize **A** in $\min_{\mathbf{A}} \sum_{i=1}^{n} \sum_{j \neq i} d(a_i, a_j)$ by block gradient descent
(5) Train k linear SVMs to update $w_i, \forall i = 1, \ldots, k$, using a_j^i as the label for feature h_j and the ith attribute, $\forall j = 1, \ldots, n, i = 1, \ldots, k$
(6) Update **A**: $a_j^i = \text{sign}(w_i' \cdot h_j), \forall j = 1, \ldots, n, i = 1, \ldots, k$
(7) **until** Convergence
(8) Return **A**, **W**

Algorithm 2: Learning visual attributes and model parameters.

It is difficult to construct effective model between low-level visual features and high-level events because of the semantic gap between them; thus we utilize visual attributes as the intermediary such that the models between attributes and features or events can be more effective. In fact, two principles proposed above reflect the property that the visual attributes should have as the intermediary.

More specifically, we propose to use max margin models for linear support vector machines (SVM) such that the attributes can be reliably predicted from the original low-level visual features. And we require the distance between the attribute representations of different events to be large; thus enough margin between events can be provided by the visual attributes. Thus, the objective function for learning visual attributes and the model parameters **W** can be written as follows:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{A}} \quad & \sum_{i=1}^{k} \|w_i\|_F^2 + C \sum_{j=1}^{n} \sum_{i=1}^{k} \xi_j^i - \frac{\gamma}{2} \sum_{i=1}^{n} \sum_{j \neq i} d\left(a_i, a_j\right) \\ \text{s.t.} \quad & \xi_j^i \geq 0 \quad \forall j = 1, \ldots, n, \ i = 1, \ldots, k \\ & a_j^i \left(w_i' \cdot h_j\right) \geq 1 - \xi_j^i \quad \forall j = 1, \ldots, n, \ i = 1, \ldots, k \\ & a_j^i = g\left(w_i' \cdot h_j\right) = \text{sign}\left(w_i' \cdot h_j\right) \\ & \forall j = 1, \ldots, n, \quad i = 1, \ldots, k. \end{aligned} \tag{24}$$

As mentioned above, we choose sign function as the activation function such that we can obtain binary visual attributes. And because each event is related to only one feature and vice versa, we just need the distance between all attribute representations to be as large as possible, leading to large margin between events. And $d(x, y)$ is the distance measure for two binary vectors x and y. Typically, Hamming distance, which counts the number of different bits between x and y, is utilized as the distance measure. w_i is the weight vector corresponding to the ith visual attribute, ξ_j^i is the slack variable corresponding to the ith attribute for the jth feature (event), and C and γ are the model parameters to balance the weight of large margin between attributes, predictability, and large margin between events.

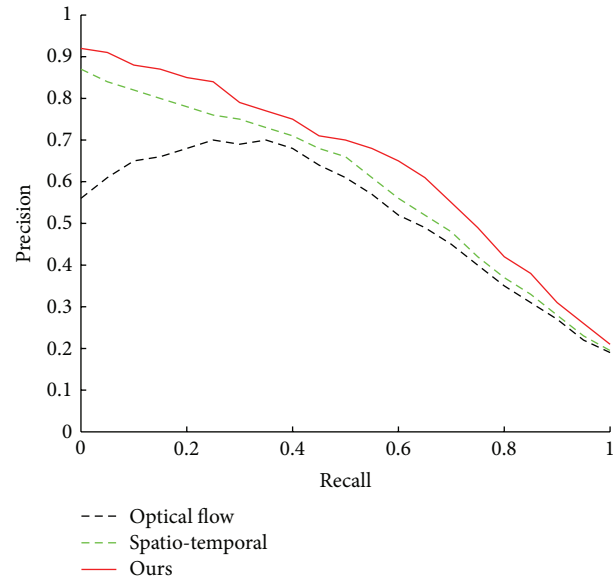

Figure 5: Performance curves on Belleview dataset.

This optimization problem is quite difficult to reach global minimum. But fortunately, local minimum can result in satisfactory performance. Here we propose to adopt an iterative strategy for the optimization problem consisting of three steps. At first, we *adjust* the value of **A** while keeping **W** fixed. The purpose of this step is to improve the margin between events represented by visual attributes, that is, **A**. Because **A** is binary, we can use block gradient descent for this step. Secondly, we update the model parameters **W** while fixing **A**. Actually, when **A** is fixed, the optimization problem for **W** can be regarded as learning k independent support vector machines. At the third step, we *update* **A** by the input low-level visual feature **H** and model parameter **W** such that the learned model parameters can indeed generate correct prediction. Here we need to point out the difference between the first and the third steps. The first step is to construct more discriminative attributes while the third step aims to make them predictable. They are quite different, but both are important. The learning algorithm is summarized in Algorithm 2.

Figure 6: Experiments on Boat-Holborn dataset.

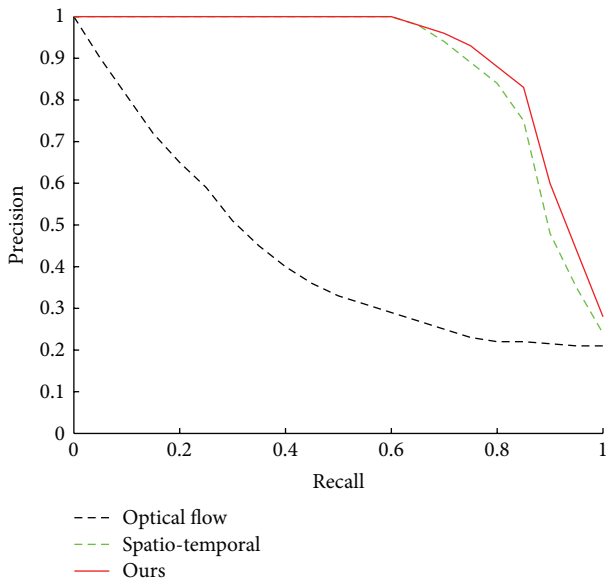

Figure 7: Performance curves on Boat-River dataset.

Actually, this alternative strategy is different from the one we proposed in the last subsection in the following perspectives. Firstly, the alternative one explicitly considers the predictability and the discriminability of the attributes and forces the model parameter to generate this kind of attributes, while the other one ignores this. Thus, intuitively, it can achieve better performance because its attributes and model are more effective. Secondly, the learning algorithm is more time consuming, especially in the first step (line 4), such that it may need offline learning while the other one just needs online learning because it is quite extremely efficient. Thirdly, the attribute learning is too precise, such that its generalization ability is weak. Therefore the background change may severely degrade the performance.

Consequently, the proposed alternative can be applied to the scenarios where the background change is slow such that the learned attributes and model can perform steadily in long time. Under such scenario, this alternative can achieve better performance. But we also need to point out that this alternative can only be applied to some specific scenarios such as indoor surveillance system where lighting conditions rarely change.

3.5. Online Update. Because the background appearance and motion are always changing in the surveillance video, such as lighting condition and weather, an anomaly detection system is expected to be self-adaptive to these changes, both in appearance and motion. Fortunately, the online update can be very efficient because of the ELM. Because the model parameters $\mathbf{W}$ and $\mathbf{B}$ are randomized, we actually do not need to change them. We just need to adjust β. In fact, the background change is usually slow during a short time (e.g., one second). So we just need to update β every second as follows.

In every second, we can randomly select some normal STVs (e.g., 10,000 to 20,000). We can obtain the low-level visual features of them and their relationship with each event. Here we need to adjust the relationship matrix $\mathbf{E}$ first because it is the output of the ELM; thus all of its elements are not strictly 1 or -1. For each STV j, we can get the most related event as $i' = \arg\max_i e_{ij}$. Typically, we have $e_{i'j} \approx 1$ but $e_{i'j} \neq 1$. So we can adjust $e_{\cdot j}$ as

$$e_{ij} = \begin{cases} 1, & \text{if } i = i' \\ -1, & \text{otherwise.} \end{cases} \tag{25}$$

Now we can get other training data $\mathbf{H}$ and $\mathbf{E}$ and we just need to retrain the ELM to update β given $\mathbf{W}$ and $\mathbf{B}$. This procedure can be quite efficient because only simple linear operations are required; thus the update can be applied to online and real-time scenarios.

4. Spatiotemporal Pyramid

Because of the presence of noise in surveillance video which is common in real-world applications, the judgement given by approach proposed above is sometimes wrong such that the detection performance is unsatisfactory for real-world application. Without postprocessing step, the detection may have low true positive rate but high false positive rate; that is, some anomalies are ignored while some normal STVs are misjudged as anomalous, which is unexpected for an effective anomaly detection approach where high true positive rate and low false positive rate are required. Fortunately, it is easy to observe that an anomalous event shows continuity in space and time; that is, it is always related to different parts in the camera and it may last for a period of time.

Figure 8: Experiments on Train dataset.

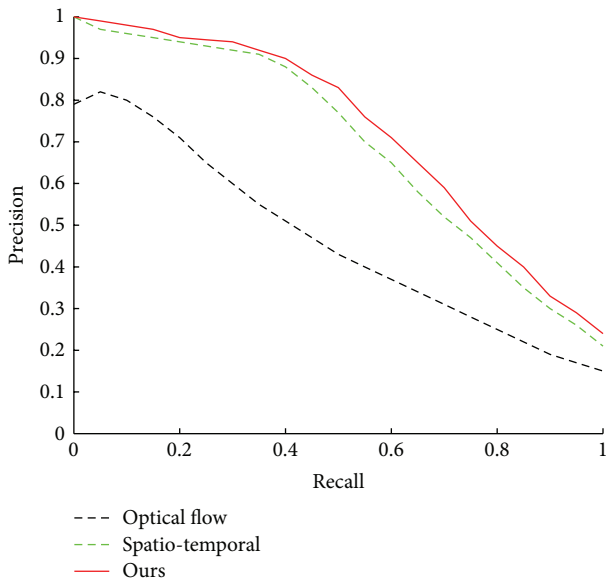

Figure 9: Performance curves on Train dataset.

Therefore, taking the spatial and temporal relationship of STVs into consideration can lead to more robust detection and significantly promote the detection performance. In this paper, we propose to use spatiotemporal pyramid (STP) to capture the relationship, as illustrated in Figure 1. We can use any levels based on the specific situation. In this paper we just use two-level STP, but we find that satisfactory result can be achieved under this setting.

Here we need to point out again that the HOG feature of upper level STV can be efficiently constructed from its lower-level STVs; thus the efficiency can be guaranteed. Actually, we can observe that the upper level STV can capture the spatial and temporal relationship of lower level STVs and the global information of an event. And because STVs in different levels of STP have different scales, our approach can perform multiscale detection based on STP. Given a STV in any scale (upper level or lower level), the anomaly judgement can be made by extreme leaning machine introduced above individually; that is, we have one ELM for each scale. Then we can combine the judgement for STVs in both upper and lower levels of STP and the judgement for neighbor STVs to give the finally refined judgement.

Before we give the judgement rules based on STP, there is an interesting and important phenomenon we need to mention, which is also essential for achieving satisfactory detection performance. The judgement on upper level STV tends to have high precision but low recall, implying that our approach can highly confidently claim that an upper level STV is anomalous, but some anomalous STVs may be missed. This is reasonable because the upper level STV can capture the global information of an event where the spatial and temporal continuity of an event can be adequately taken into consideration while some important local details will be ignored. On the other hand, the judgement on the lower level STV usually has low precision but high recall; that is, it tends to treat a STV as anomalous because it is too sensitive to local details and noise and ignores the spatial and temporal relationship between STVs, but it can capture more local information than upper level STV. Thus we propose the spatiotemporal pyramid to combine these two-level STVs to take advantage of both local details and global information simultaneously as follows.

On one hand, if it is judged to be abnormal for an upper STV, this judgement is highly confidential. If it is judged to be normal, it still has marked probabilistic to be anomalous. Thus the following results should be take into consideration too: (1) its six neighbors which consider the spatial and temporal continuity of events and (2) its lower STVs which capture the local detail information of events. To utilize all these pieces of information above, in this paper, an upper STV is finally judged to be anomalous if any of the following three criteria is satisfied: (1) it is judged to be anomalous, (2) at least three of its neighbors are anomalous, and (3) at least five of its lower level STVs are anomalous. Actually, the high-precision result for upper level STV leads to the first criterion. The second criterion is based on the spatial and temporal continuity of events. And the third criterion is based on a voting scheme because it is reasonable to assume that though one lower STV may be influenced by noise or local details, it is difficult for most STVs to generate wrong judgement.

On the other hand, lower-level STVs cannot capture the spatial and temporal continuity of events and they are significantly affected by noise in surveillance videos; consequently the judgement of a STV should be incorporated with its upper level STV and neighbors which consider the global information and the spatial temporal continuity of STVs. Therefore a lower-level STV is finally considered to be

Figure 10: Experiments on Ped1 dataset.

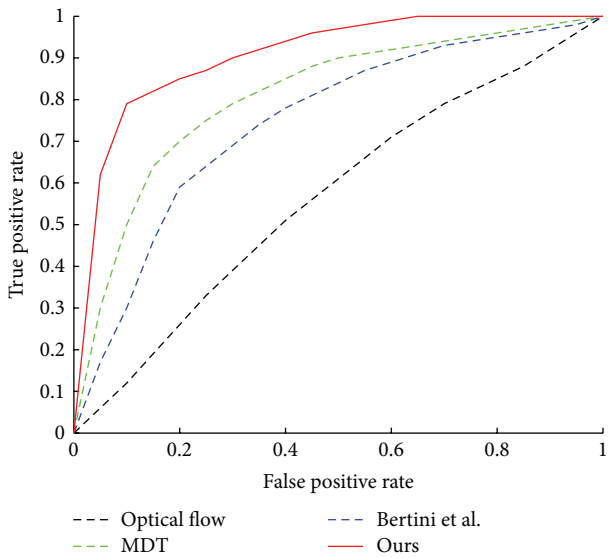

Figure 11: Performance curves on Ped1 dataset.

anomalous if it is 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 spatiotemporal pyramid and criteria above, we take into consideration the spatial and temporal continuity of events, the relationship between STVs in space and time, and the local details simultaneously which can significantly promote the performance. Furthermore, the spatiotemporal pyramid allows us to perform multiscale detection because STVs in different levels have different scales.

5. Experiment

To verify the effectiveness of the proposed approach, we test it in the following two publicly available datasets for anomaly detection: anomaly behavior detection dataset [53] (http://www.cse.yorku.ca/vision/research/) and UCSD pedestrian dataset [18] (http://www.svcl.ucsd.edu/projects/anomaly). The evaluation and comparison of different approaches are presented in two kinds of performance curves, that is, precision-recall, ROC curves, and equal error rate (EER) at both frame level and pixel level is also reported. As mentioned above, we use a two-level pyramid, and 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 the number of visual attributes, that is, the number of hidden units in extreme learning machine, to 256, and the anomaly probabilistic $p_a = 10^{-3}$. In fact, our method does not require any training data because it is totally unsupervised. It just use the first one or two seconds of a test video to construct the initial normal events set. Furthermore, we set that the extreme learning machine is updated every one second. Furthermore, the following several state-of-the-art approaches for anomaly detection are compared to our approach: optical flow [15], Mahadevan et al. [18], sparse reconstruction (Cong et al.) [19], Zaharescu and Wildes [53], Reddy et al. [52], and Bertini et al. [20].

The first dataset is *Belleview* dataset. It is a traffic scene where the lighting conditions change during the day gradually. Cars running from top to bottom or vice versa is normal event, while cars entering or exiting from the intersection from left or right and people in the lane are the anomalous events. The second is *Boat-River* dataset. The normal events are the waves in the river, while the anomalous event is defined as a boat that passing the scene. Actually, the boat is the newly observed object in the scene. The third is *Train* dataset. The normal events are people sitting on the train and some background change while anomalies are moving people. This dataset is quite challenging because the illumination varies drastically and the camera is not stable. 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 Figures 4, 5, 6, 7, 8, and 9, respectively. We can observe that our approach is superior to state-of-the-art methods, for example, Zaharescu and Wildes. The superiority of our approach is based on the following three main reasons: (1) our approach address the semantic gap between low-level visual features and high-level events via visual attributes as intermediary such that more effective model can be constructed between them, (2) our approach can update model frequently (every one second) and quite efficiently such that it is very robust to drastic background change, (3) the spatial and temporal relationship between neighbor STVs is fully taken into consideration in our approach; therefore it is robust to local noise, and (4) our approach considers the spatial and temporal continuity of anomalous event; thus complicated events which may last several seconds or cover a large region can be effectively detected.

The UCSD dataset contains two subsets corresponding to two different scenes captured by fixed cameras overlooking the pedestrian walkways: Ped1 in which people with some distortion move towards and away from the camera and

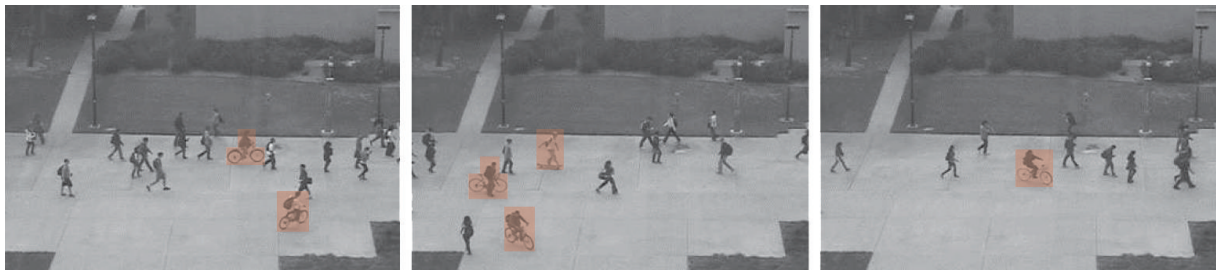

Figure 12: Experiments on Ped2 dataset.

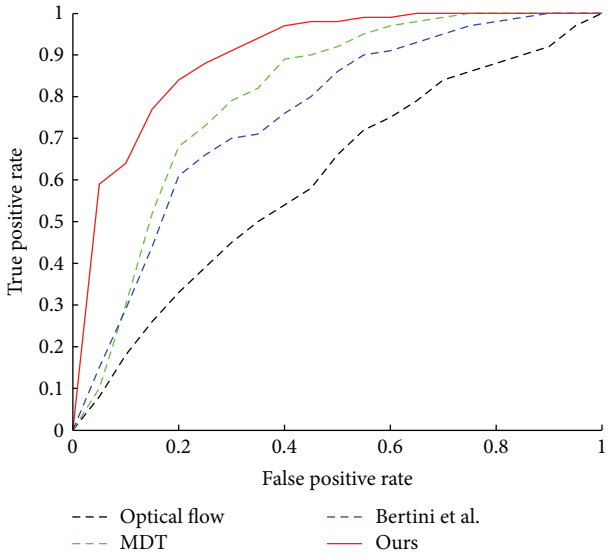

Figure 13: Performance curves on Ped2 dataset.

Ped2 which shows the pedestrian movement parallel to the camera. All videos are recorded with 10 frames per second. The resolution of Ped1 and Ped2 is 238 × 158 and 360 × 240, respectively. Specifically, pedestrians walking in walkways are regarded as normal events, while nonpedestrian objects, for example, small carts, or nonwalking pedestrians, for example, cyclists and skaters, are treated as the anomalies. We test our approach on both subsets. We follow the evaluation adopted in [18, 20]. 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. Because of the "lucky guess" which means the detected anomalies are different from the true anomalies in a frame, the frame level evaluation is sometimes not convincing enough because it does not take this phenomenon into consideration. Thus we adopt pixel level evaluation. We report the anomalous regions detected by our approach, the ROC curves, and the equal error rate (EER) which is the rate where the false positive rate is equal to 1 minus the true positive rate. The anomalous regions detected by our approach, the ROC curves, and the ERR of different approaches are shown in Figures 10, 11, 12, and 13 and Table 1, respectively. From the experiment results, we can observe that our approach can significantly outperform the state-of-the-art approaches on both subsets for both frame level detection and pixel level detection, especially compared to real-time detection approaches, which validates the effectiveness of our approach for real-time anomaly detection in surveillance videos.

Table 1: Comparison of the proposed approach and the state-of-the-art approaches 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* [15]	38%	76%	42%	80%
Mahadevan et al. [18]	25%	58%	25%	55%
Cong et al. [19]	19%	—	20%	—
Reddy et al.* [52]	22.5%	32%	21%	31%
Bertini et al.* [20]	31%	70%	30%	68%
Ours*	**17%**	**28%**	**16%**	**26%**

6. Conclusion

In this paper, a novel automatic anomaly detection approach is proposed. Densely sampled spatiotemporal video volumes represented by spatiotemporal local features are the fundamental of our approach. Normal event set is efficiently constructed from test data in an unsupervised way. We use visual attribute as intermediary to bridge the large semantic gap between low-level visual feature and high-level event. Extreme learning machine is utilized to model this three-level framework and it can be efficiently updated such that our approach is adaptive to background change. We propose to use spatiotemporal pyramid to capture the relationship between different STVs and the spatial and temporal continuity of anomalous events. Extensive experiments on several public datasets are conducted and the superior performance compared to several state-of-the-art approaches verifies the effectiveness of our approach.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] T. Troscianko, A. Holmes, J. Stillman, M. Mirmehdi, D. Wright, and A. Wilson, "What happens next? The predictability of natural behaviour viewed through CCTV cameras," *Perception*, vol. 33, no. 1, pp. 87–101, 2004.

[2] H. Keval and M. A. Sasse, "'Not the usual suspects': a study of factors reducing the effectiveness of CCTV," *Security Journal*, vol. 23, no. 2, pp. 134–154, 2010.

[3] N. Haering, P. L. Venetianer, and A. Lipton, "The evolution of video surveillance: an overview," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 279–290, 2008.

[4] C. Brax, L. Niklasson, and M. Smedberg, "Finding behavioural anomalies in public areas using video surveillance data," in *Proceedings of the 11th International Conference on Information Fusion (FUSION '08)*, pp. 1–8, Cologne, Germany, July 2008.

[5] I. Ivanov, F. Dufaux, T. M. Ha, and T. Ebrahimi, "Towards generic detection of unusual events in video surveillance," in *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS '09)*, pp. 61–66, Genoa, Italy, September 2009.

[6] P. Antonakaki, D. Kosmopoulos, and S. J. Perantonis, "Detecting abnormal human behaviour using multiple cameras," *Signal Processing*, vol. 89, no. 9, pp. 1723–1738, 2009.

[7] S. Calderara, C. Alaimo, A. Prati, and R. Cucchiara, "A real-time system for abnormal path detection," in *Proceedings of the IEEE 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP '09)*, December 2009.

[8] C. Liu, G. Wang, W. Ning, X. Lin, L. Li, and Z. Liu, "Anomaly detection in surveillance video using motion direction statistics," in *Proceedings of the 17th IEEE International Conference on Image Processing (ICIP '10)*, pp. 717–720, IEEE, Hong Kong, September 2010.

[9] C. Piciarelli and G. L. Foresti, "Surveillance-oriented event detection in video streams," *IEEE Intelligent Systems*, vol. 26, no. 3, pp. 32–41, 2011.

[10] C. C. Loy, T. Xiang, and S. Gong, "Detecting and discriminating behavioural anomalies," *Pattern Recognition*, vol. 44, no. 1, pp. 117–132, 2011.

[11] R. R. Sillito and R. B. Fisher, "Semi-supervised learning for anomalous trajectory detection," in *Proceedings of the 19th British Machine Vision Conference (BMVC '08)*, September 2008.

[12] C. Piciarelli and G. L. Foresti, "On-line trajectory clustering for anomalous events detection," *Pattern Recognition Letters*, vol. 27, no. 15, pp. 1835–1842, 2006.

[13] O. Boiman and M. Irani, "Detecting irregularities in images and in video," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 17–31, 2007.

[14] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 893–908, 2008.

[15] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[16] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*, pp. 935–942, Miami, Fla, USA, June 2009.

[17] F. Jiang, Y. Wu, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 907–913, 2009.

[18] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '10)*, pp. 1975–1981, San Francisco, Calif, USA, June 2010.

[19] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '11)*, pp. 3449–3456, June 2011.

[20] M. Bertini, A. Del Bimbo, and L. Seidenari, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 320–329, 2012.

[21] M. J. Roshtkhari and M. D. Levine, "Online dominant and anomalous behavior detection in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '12)*, 2012.

[22] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops)CVPR '09)*, pp. 1446–1453, Miami, Fla, USA, June 2009.

[23] S. Khalid, "Activity classification and anomaly detection using *m*-mediods based modelling of motion patterns," *Pattern Recognition*, vol. 43, no. 10, pp. 3636–3647, 2010.

[24] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011.

[25] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, no. 1–3, pp. 489–501, 2006.

[26] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR '09)*, pp. 1778–1785, Miami, Fla, USA, June 2009.

[27] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *Proceedings of the 12th International Conference on Computer Vision (ICCV '09)*, pp. 365–372, IEEE, Kyoto, Japan, October 2009.

[28] M. Rastegari, A. Farhadi, and D. Forsyth, "Attribute discovery via predictable discriminative binary codes," in *Computer Vision—ECCV 2012: Proceedings of the 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Part VI*, vol. 7577 of *Lecture Notes in Computer Science*, pp. 876–889, Springer, Berlin, Germany, 2012.

[29] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06)*, vol. 2, pp. 2169–2178, June 2006.

[30] B. T. Morris and M. M. Trivedi, "Trajectory learning for activity understanding: unsupervised, multilevel, and long-term adaptive approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2287–2301, 2011.

[31] K. Ouivirach, S. Gharti, and M. N. Dailey, "Incremental behavior modeling and suspicious activity detection," *Pattern Recognition*, vol. 46, no. 3, pp. 671–680, 2013.

[32] Y. Benezeth, P.-M. Jodoin, and V. Saligrama, "Abnormality detection using low-level co-occurring events," *Pattern Recognition Letters*, vol. 32, no. 3, pp. 423–431, 2011.

[33] T. Hospedales, S. Gong, and T. Xiang, "Video behaviour mining using a dynamic topic model," *International Journal of Computer Vision*, vol. 98, no. 3, pp. 303–323, 2012.

[34] Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal events detection based on spatio-temporal co-occurences," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops '09)*, pp. 2458–2465, Miami, Fla, USA, June 2009.

[35] E. B. Ermis, V. Saligrama, P.-M. Jodoin, and J. Konrad, "Motion segmentation and abnormal behavior detection via behavior clustering," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '08)*, pp. 769–772, IEEE, San Diego, Calif, USA, October 2008.

[36] K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis, "Real-time foreground-background segmentation using codebook model," *Real-Time Imaging*, vol. 11, no. 3, pp. 172–185, 2005.

[37] A. Mittal, A. Monnet, and N. Paragios, "Scene modeling and change detection in dynamic scenes: a subspace approach," *Computer Vision and Image Understanding*, vol. 113, no. 1, pp. 63–79, 2009.

[38] X. Zhu and Z. Liu, "Human behavior clustering for anomaly detection," *Frontiers of Computer Science in China*, vol. 5, no. 3, pp. 279–289, 2011.

[39] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

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

[41] J. Li, S. Gong, and T. Xiang, "Learning behavioural context," *International Journal of Computer Vision*, vol. 97, no. 3, pp. 276–304, 2012.

[42] E. Ricci, G. Zen, N. Sebe, and S. Messelodi, "A prototype learning framework using EMD: application to complex scenes analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 513–526, 2012.

[43] P.-M. Jodoin, J. Konrad, and V. Saligrama, "Modeling background activity for behavior subtraction," in *Proceedings of the 2nd ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC '08)*, pp. 1–10, Stanford, Calif, USA, September 2008.

[44] P. Jodoin, V. Saligrama, and J. Konrad, "Behavior subtraction," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4244–4255, 2012.

[45] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM International Conference on Multimedia (MULTIMEDIA '07)*, pp. 357–360, September 2007.

[46] Y. Nesterov, "Gradient methods for minimizing composite objective function," CORE Discussion Papers no. 2007-76, CORE, 2007.

[47] J. W. Cao, T. Chen, and J. Fan, "Fast online learning algorithm for landmark recognition based on bow framework," in *Proceedings of the 9th IEEE Conference on Industrial Electronics and Applications*, June 2014.

[48] J. W. Cao and Z. Lin, "Bayesian signal detection with compressed measurements," *Information Sciences*, vol. 289, pp. 241–253, 2014.

[49] Y. Jin, J. Cao, Q. Ruan, and X. Wang, "Cross-modality 2D-3D face recognition via multiview smooth discriminant analysis based on ELM," *Journal of Electrical and Computer Engineering*, vol. 2014, Article ID 584241, 9 pages, 2014.

[50] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Applications*, Wiley-Interscience, 1971.

[51] D. Serre, *Matrices: Theory and applications*, Springer, 2002.

[52] V. Reddy, C. Sanderson, and B. C. Lovell, "Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '11)*, pp. 55–61, Colorado Springs, Colo, USA, June 2011.

[53] A. Zaharescu and R. Wildes, "Anomalous behaviour detection using spatiotemporal oriented energies, subset inclusion histogram comparison and event-driven processing," in *Computer Vision—ECCV 2010*, pp. 563–576, 2010.