Unsupervised Abnormal Crowd Activity Detection Using Semiparametric Scan Statistic

Yang Hu, Yangmuzi Zhang, Larry S. Davis
Institute for Advanced Computer Studies, University of Maryland, College Park, MD 20742
{yanghu, ymzhang, lsd}@umiacs.umd.edu

Abstract

We propose a fully unsupervised method for abnormal activity detection in crowded scenes. Neither normal nor abnormal training examples are needed before detection. By observing that in crowded scenes, normal activities are the behaviors performed by the majority of people and abnormalities are behaviors that occur rarely and are different from most others, we propose to use a scan statistic method to solve the problem. It scans a video with windows of variable shape and size. The abnormality of each window is measured by a likelihood ratio test statistic, which compares two hypotheses about whether or not the characteristics of the observations inside and outside the window are different. A semiparametric density ratio method is used to model the observations, which is applicable to a wide variety of data. To reduce the search complexity of the sliding window based scanning, a fast two-round scanning algorithm is proposed. We successfully applied our algorithm to detect activities that are anomalous in different ways, achieving performance competitive to other state-of-the-art methods which requiring supervision.

1. Introduction

Identifying abnormal activities in densely crowded scenes has been attracting increasing attention in the computer vision community. This problem plays an important role in many applications such as crowd surveillance, public place monitoring, security control, etc. The main paradigm in this field is to assume the availability of a set of normal examples before detection, which define what normal activities look like. Abnormality of a new observation is then measured either by its similarity with the given examples or by its compatibility with the model derived from the examples.

The requirement of normal activity being defined beforehand may complicate the deployment of the approaches in real applications. For example, an important characteristics of normal and abnormal activities in practice is their relativity. Abnormal activities in one situation may become normal in others. Fig. 1(b) shows a scene in a subway station. The gate is used as an exit gate and people entering it are regarded as abnormal. However, at other times the same gate may be changed to an entrance gate and people entering it becomes normal while people exiting through it becomes abnormal. To apply detection approaches based on the above paradigm, one must know in advance whether the gate is serving as an entrance or an exit gate at a specific time, so that the correct normal activity model can be used.

(a) (b) (c)

Figure 1. Examples of abnormal activities in crowded scenes. (a) Global abnormal activity of crowd panic. (b) Local abnormal activity where a man enters through the exit gate in a subway station. (c) Local abnormal activity in a highly dynamic scene where a man steps out while the crowd is dancing.

To reduce the requirement for external examples of normal activity, some recent work assumes that a certain duration in the beginning of a video contains only normal behaviors, which are used to train the normal activity model [12, 5, 22]. However, these methods cannot be applied when the normal behaviors keep changing over time so that the behaviors in the beginning of a video cannot be used to infer what is normal in the remainder of the video. They are also not applicable when the abnormal behaviors appear right after the video begins. The flash mob video shown in Fig. 1(c) is such an example. The normal behavior, which is the dancing of the crowd, changes quickly. And the abnormal behavior of a man stepping out of the crowd happens near the beginning of the video.

To address the above limitations of previous approaches, we propose a fully unsupervised solution for the problem, which requires neither normal nor abnormal examples being provided beforehand. In crowded scenes, normal activities are the behaviors performed by the majority of people and abnormal activities are behaviors that occur rarely and are different from most others in the scene. We therefore propose to use a scan statistic method to solve the problem. The main idea is to scan the video using a large number of windows with variable shape and size. For each window, two hypotheses about whether or not the characteristics of the observations inside the window is different from those outside it are compared. A likelihood ratio test statistic, which is based on a semiparametric density ratio model, is computed for the comparison and used as a measure of the window's abnormality.

Besides the waiver of the requirement for normal activity examples, the novelty and advantages of the scan statistic method also lie in the following aspects. First, as a quite general detection framework, it can be used to detect different kinds of abnormal events, including: global abnormal activities, where the whole crowd is involved in the abnormal event in which their behaviors are different from those at most other times in the video (Fig. 1(a)); local abnormal activities, where the behaviors of a small group of people are different from those of most others in the video (Fig. 1(b)). It can also be applied to videos in which the normal behaviors are highly dynamic as in the flash mob example(Fig. 1(c)). Second, in previous scanning window based approaches [5, 22], the shape and size of the local windows are usually fixed and the abnormalities are measured independent of each other. In our scan statistic method, windows with variable shapes and sizes are used so that an abnormal event can be captured in a single window instead of being divided up, which allows its statistical characteristics to be measured more accurately. Third, a semiparametric density ratio method [11] is used to model the observations inside and outside a scanning window. This reduces the requirement for assuming a specific parametric probability model and makes it directly applicable to different types of observations. Forth, to reduce the computational complexity of exhaustive search, we present a fast scanning algorithm for the semiparametric scan statistic method. It dramatically reduces the number of regions being examined while still maintaining good accuracy.

Scan statistics are a powerful method for cluster detection. They have applications in many fields, such as epidemiology, criminology, genetics, mining, astronomy, etc. However, their usage for solving computer vision problems has not been exploited before. In this paper, we illustrate its application for abnormal activity detection in crowded scenes. Experiments on both benchmark datasets and videos "in the wild" validate the effectiveness of the method.

2. Related Work

Various approaches have been proposed for anomaly detection, for both crowded and non-crowded scenes. They can be broadly categorized according to whether or not examples of normal and abnormal activities are needed before detection. The first type of work treats the task as a binary classification problem. To train the classifier, not only normal but also abnormal activities are needed. For example, in [16, 6], both activities were used to train a support vector machine (SVM) for abnormal activity recognition.

Considering the rich patterns of irregular behaviors, the second type of work finds abnormalities without knowing what they are beforehand. Although not requiring abnormal examples, training data are still needed to define what normal activities look like. Note that since abnormal examples are not used during training, some approaches of this type have been claimed to be unsupervised [4, 22, 21, 2]. According to how normal examples are used, the methods can be further divided into two sub-categories. The first one contains data-driven methods. They directly compare new observations with a set of normal examples. The observations whose similarity scores are low [4, 2] or which cannot be composed well by the known examples [3] are regarded as abnormal. In the second sub-category, normal examples are used to build explicit models for normal activities. An anomaly is declared when the new observation cannot be explained well by the model. Kim and Grauman [12] used Mixture of Probabilistic PCA to model normal local activity patterns. In [17, 8], Latent Dirichlet Allocation (LDA) was used to discover latent topics in normal activities. With the recent popularity of sparse coding, normal basis sets are also learned from normal activities, over which sparse reconstruction costs are computed for new observations [22, 5].

Our work belongs to the third type where neither normal nor abnormal examples are required before detection. Several such approaches have been presented, mainly for non-crowded scenes [23, 3, 19, 10]. The main idea of both [23] and [3] is to use the input video itself to build the reference database and compare each event with all other events observed. However, for crowded scenes, performing such pairwise comparison is time-consuming. Our method collectively models the whole crowd's behaviors, which is more appropriate and efficient for crowded scenes. Although in [10] Ito et al. also used the density ratio estimation technique, their method cannot locate the spatial-temporal regions of local abnormal events.

Among recent work that focuses on crowded scenes, much effort has been devoted to design descriptive features to characterize crowded scenes, such as social force model [17], chaotic invariants of particle trajectories [20] and mixture of dynamic textures [15]. Instead of designing a new video representation, we propose a new detection framework which can be combined with different representations. We show that competitive performance can be achieved even only using optical flow as the descriptors.

3. Anomaly Detection with Scan Statistics

3.1. The Scan Statistic Method

To discover spatial-temporal regions in which activities are abnormal, we scan the video with a three-dimensional sliding window. Both the shape and size of the window are variable so that abnormal activities that occupy irregular regions in the video can be identified and localized.

For a region S defined by a window, we measure whether the observations in it is anomalous. Specifically, letting the video be our whole study region and $\mathbf{x}$ be the set of observations in it, we define two hypotheses about $\mathbf{x}$. The null hypothesis H_0 assumes no anomaly exists in $\mathbf{x}$, i.e. the underlying characteristics of the observations throughout the whole study region are the same. The alternative hypothesis $H_1(S)$ assumes that the characteristics of observations inside and outside S are different. These two hypotheses can be compared by computing the likelihood ratio test statistic:

$$\lambda(S) = \frac{\Pr(\mathbf{x}|H_1(S))}{\Pr(\mathbf{x}|H_0)} \tag{1}$$

where $\Pr(\mathbf{x}|H_i)$ is the likelihood function of observing $\mathbf{x}$ under hypothesis $H_i, i = 0, 1$.

The likelihood ratio test is a measure of the strength of H_1. The larger this number is, the more likely H_1 is true and the greater the discrepancy between distributions of the observations inside and outside S. Therefore we can use $\lambda(S)$ as a measure of the abnormality of the region.

3.2. Semiparametric Density Ratio Model

To compute the likelihood ratio test statistic in Eq.(1) for each region, an appropriate probability model must be postulated for $\Pr(\mathbf{x}|H_i)$. Instead of assuming a specific parametric probability model as in Kulldorff's method [13], we follow Kedem and Wen [11] and use a semiparametric density ratio method to compute the likelihood for the observations under different hypotheses.

3.2.1 The Density Ratio Model

A certain scanning window separates the whole study region into two parts. This results in two sample sets,

$$\begin{aligned} \mathbf{x}_1 &= (x_{11}, x_{12}, \ldots, x_{1n_1})^T \sim f(x) \\ \mathbf{x}_2 &= (x_{21}, x_{22}, \ldots, x_{2n_2})^T \sim g(x) \end{aligned} \tag{2}$$

where $\mathbf{x}_1$ contains the observations inside the window with sample size n_1, and $\mathbf{x}_2$ represents the observations outside the window with sample size n_2. $f(x)$ and $g(x)$ are their corresponding probability density functions. Taking $g(x)$ as the reference density, we assume that the ratio between $f(x)$ and $g(x)$ has an exponential form

$$\frac{f(x)}{g(x)} = \exp(\alpha + \boldsymbol{\beta}^T \mathbf{h}(x)) \tag{3}$$

Here $\mathbf{h}(x)$ is a predefined function of x which may take a form as x, x^2, $\log(x)$, or any combination of them such as $(x, x^2)^T$ or $(x, \log(x))^T$. α is a scalar, $\boldsymbol{\beta}$ could be a scalar or vector depending on $\mathbf{h}(x)$.

It is obvious that $\boldsymbol{\beta} = 0$ implies $\alpha = 0$ and therefore we have $f(x) \equiv g(x)$, which means the samples inside and outside the scanning window come from a common distribution. Therefore, the null hypothesis of no anomaly in the study region can be represented as $H_0 : \boldsymbol{\beta} = 0$.

Let $n = n_1 + n_2$, $\mathbf{t} = (t_1, t_2, \ldots, t_n)^T = (x_{11}, \ldots, x_{1n_1}, x_{21}, \ldots, x_{2n_2})^T$ denotes the combined sample set, and let $p_i = dG(t_i) = \Pr(X = t_i)$, where $G(x)$ is the corresponding cumulative distribution function of $g(x)$. The semiparametric likelihood of the samples can be represented as

$$L(\alpha, \boldsymbol{\beta}, G) = \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp(\alpha + \boldsymbol{\beta}^T \mathbf{h}(x_{1j})) \tag{4}$$

The likelihood ratio score in Eq.(1) are then computed by estimating this likelihood with and without the assumption of $f(x) \equiv g(x)$.

3.2.2 Parameter Estimation

We can follow the profiling procedure discussed in [18, 7] to estimate $p_i, i = 1, \ldots, n$ and $(\alpha, \boldsymbol{\beta})$. It first fixes $(\alpha, \boldsymbol{\beta})$ and maximize L in Eq.(4) with respect to $p_i, i = 1, \ldots, n$ subject to constraints $\sum p_i = 1$ and $\sum p_i[\omega(t_i) - 1] = 0$, where $\omega(x) = \exp(\alpha + \boldsymbol{\beta}^T \mathbf{h}(x))$ and the constraints reflect the fact that $dG(x)$ and $\omega(x)dG(x)$ are both distribution functions. By employing the Lagrange multiplier method, we are able to represent p_i in terms of α and $\boldsymbol{\beta}$. It then substitutes p_i back into the log-likelihood $\ell = \log L(\alpha, \boldsymbol{\beta}, G)$. By setting $\partial\ell/\partial\alpha = 0$, we have that the maximum value of L is attained at $p_i = n_2^{-1}[1 + \rho\omega(t_i)]^{-1}$, where $\rho = \frac{n_1}{n_2}$. Therefore, ignoring a constant, the log-likelihood as a function of $\alpha, \boldsymbol{\beta}$ only is

$$\begin{aligned} \ell(\alpha, \boldsymbol{\beta}) = &- \sum_{i=1}^{n} \log[1 + \rho \exp(\alpha + \boldsymbol{\beta}^T \mathbf{h}(t_i))] \\ &+ \sum_{j=1}^{n_1} [\alpha + \boldsymbol{\beta}^T \mathbf{h}(x_{1j})] \end{aligned} \tag{5}$$

α and $\boldsymbol{\beta}$ are then estimated by maximizing this log-likelihood through a Newton's method.

3.2.3 Computing the Likelihood Ratio Test Statistic

Let $\hat{\alpha}, \hat{\boldsymbol{\beta}}$ be the maximum likelihood estimates of α and $\boldsymbol{\beta}$; then $\ell(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ corresponds to the log-likelihood of observing the samples without the assumption of $f(x) \equiv g(x)$, i.e. under the alternative hypothesis $H_1(S)$. $\ell(0, \mathbf{0})$ corresponds to the log-likelihood under the null hypothesis $H_0 : \boldsymbol{\beta} = 0$. The likelihood ratio test statistic $\lambda(S)$ in Eq.(1) is twice the difference in these log-likelihoods:

$$\begin{aligned}\lambda(S) \equiv & -2[\ell(0, \mathbf{0}) - \ell(\hat{\alpha}, \hat{\boldsymbol{\beta}})] \qquad (6)\\ = & -2\sum_{i=1}^{n} \log[1 + \rho \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{h}(t_i))] \\ & + 2\sum_{j=1}^{n_1} [\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \mathbf{h}(x_{1j})] + 2n \log(1 + \rho)\end{aligned}$$

This semiparametric density ratio method has two major advantages. First, the parameters $\alpha, \boldsymbol{\beta}$ and the distributions are estimated from the combined data $\mathbf{t}$, not just from samples either inside or outside the scanning window. When the window size is small, estimating the two distributions separately will be unreliable since the number of samples inside the window is too small. Second, the method does not require assumptions about the specific parametric probability model of the data. Although we still have to chose the form of $\mathbf{h}(x)$, it is not specific for different distributions. For example, it is easy to derive that for both Bernoulli and Poisson distributions, we have $\mathbf{h}(x) = x$. Therefore, the same method can be directly applied to problems with quite different data distributions.

4. A Fast Scanning Algorithm

One major challenge of the scan statistic method is the large number of regions that need to be scanned. Due to the highly dynamic nature of human motion, the abnormal activities usually occupy irregular regions in the three-dimensional space. This requires the scanning window to cover regions with a large variety of shapes and sizes. Moreover, there are usually multiple anomalies in the video. This makes fast detection methods that only search for the most anomalous region inapplicable. In this section, we propose an efficient search scheme for the semiparametric scan statistic method.

4.1. Motion Representation

We use optical flow as a low-level measure of motion in videos. The optical flow field for each frame is estimated using the method in [14]. We then build a bag of words representation for the flow field. We define a dense grid of rectangular cells on the image. In each cell we extract a local patch of optical flow. The flow patch is then quantized to a flow word. Therefore, the samples for the scan statistic method take the values of $t_i = 1, \ldots, K$, where K is the size of the code book. Then, the motion in a spatial-temporal region is represented by the histogram of flow words.

The histogram representation simplifies the computation of the log-likelihood $\ell(\alpha, \boldsymbol{\beta})$ in Eq.(5). Let $\mathbf{b} = (b_1, b_2, \ldots, b_K)^T$ be the histogram for a spatial-temporal region, where $b_k, k = 1, \ldots, K$ is the number of flow patches that are quantized to the kth word in the region. Letting $\mathbf{b}^w = (b_1^w, b_2^w, \ldots, b_K^w)^T$ be the histogram of the whole study region, Eq.(5) can be rewritten as

$$\begin{aligned}\ell(\alpha, \boldsymbol{\beta}) = & -\sum_{k=1}^{K} b_k^w \log[1 + \rho \exp(\alpha + \boldsymbol{\beta}^T \mathbf{h}(k))] \qquad (7)\\ & + \sum_{k=1}^{K} b_k [\alpha + \boldsymbol{\beta}^T \mathbf{h}(k)]\end{aligned}$$

$\mathbf{h}(k)$ can be pre-computed here instead of computing it for each sample separately as in Eq.(5). We set $\mathbf{h}(x) = (x, \log(x))^T$ in the experiments. The simplification of the log-likelihood in Eq.(7) makes the estimation of $\alpha, \boldsymbol{\beta}$ for each region very efficient. The computation of the likelihood ratio test statistic in Eq.(6) is simplified similarly.

4.2. A Two-round Scanning Algorithm

Since an irregularly shaped spatial-temporal region can be approximated by a union of a set of equally sized cuboids, instead of searching exhaustively with windows of variable shape and size, we first scan the video with a fixed-sized three-dimensional rectangular window. Note that by using the integral image trick, the flow word histogram for any cuboid in the video can be obtained very efficiently. According to the definition of the likelihood ratio test, regions that contain abnormal activities should always have larger $\lambda(S)$ than those that only contain normal activities. However, due to the complexity of human motion, some regions may have large $\lambda(S)$ and look anomalous although they are part of the normal activity. We therefore run a second round of scan which not only reduces these false alarms but also helps to identify the true scope of the irregularly shaped anomalous regions.

Among the detections of the first round of scan, false anomalous regions are usually isolated. On the other hand, true anomalous regions are detected in groups. They constitute larger anomalous regions. Therefore, after the first round of search, we take regions whose $\lambda(S)$ are larger than a threshold T and build a graph over them. If two regions are adjacent to each other, an edge is added between them. Then we find the connected components in this graph. Since different T results in different graphs and so different connected components, we use multiple thresholds to explore different possibilities for connecting the regions. Specifically, let $\hat{\lambda}$ be the largest $\lambda(S)$ in the first round of scan,

Figure 2. Examples of detected abnormal activities in the UMN dataset.

$\{T_i = q_i\hat{\lambda}|q_i = 0.1, 0.2, \ldots, 0.9\}$ are used in experiments. Finally we compute the likelihood ratio test $\lambda(S)$ for the connected components. In true anomalous areas, $\lambda(S)$ of these new formed regions are larger than those of the regions that constitute them.

This is a greedy algorithm for scanning the video. The first round of scan provides initial estimation of the probabilities that any local spatial-temporal region in the video contains abnormal activities. In the second round of scan, a set of most probable candidates are examined. For a video divided into an $m \times n \times t$ grid, both rounds scan $O(mnt)$ regions. The exhaustive search scheme that uses a window of variable shape and size needs to search $O(m^2n^2t^2)$ regions even only considering regions of cubic shape. The above greedy algorithm not only examines far fewer regions, but also detects regions with irregular shapes.

5. Experimental Results

In this section, we show the performance of the scan statistic method on three benchmark datasets and several videos "in the wild" that are downloaded from the web. These datasets show examples of different kinds of global and local abnormal activities.

5.1. UMN Dataset

The UMN dataset [9] includes 11 video sequences of three different scenes of crowded escape events. Each video begins with normal behavior where people walk around, followed by a sudden abnormal panic. Some examples of the abnormal behaviors are shown in Fig. 2. Since the whole crowd is involved in the abnormal activities, we use this dataset to evaluate the performance for global abnormal event detection.

Sine we only need to detect in which frames the activities are anomalous, we fix the spatial size of the scanning window to cover the whole image. To quantitatively compare the performance of our algorithm with other state-of-the-art methods, we take a window's score $\lambda(S)$ as the score for the frame in the middle of the window. Here we only need to run the first round of scan with fixed sized window. The window length is set to 20 frames. We then draw ROC curve for this frame-level measurement. Table 1 compares the corresponding average AUC with other methods. Our scan statistic method achieved very competitive performance. Note that all other methods require examples of normal activities being provided, which are used either to train their models or as references for comparison. Our method does not have this requirement. No training or initialization are involved. The result demonstrates the effectiveness of the scan statistic method for global abnormal events detection in videos.

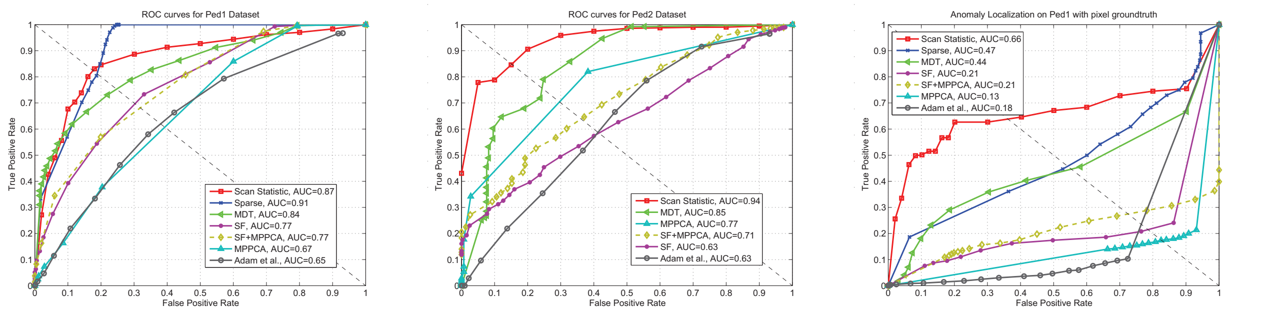
Method	Area under ROC
Chaotic invariants [20]	0.99
Social force [17]	0.96
Optical flow [17]	0.84
Nearest neighbor [5]	0.93
Sparse Reconstruction [5] (Scene1/Scene2/Scene3)	0.995/0.975/0.964
Scan statistic (Scene1/Scene2/Scene3)	0.991/0.951/0.99

Table 1. The comparison of our semiparametric scan statistic method with other methods on the UMN dataset.

5.2. UCSD Anomaly Detection Dataset

The UCSD anomaly dataset contains two subsets, Ped1 and Ped2, each corresponding to a different scene. The first scene contains groups of people walking towards and away from the camera. The second scene contains pedestrian movement parallel to the camera plane. Abnormal events mainly include non pedestrian entities, such as people riding bicycles, skaters, carts, and anomalous pedestrian motion patterns. It provides training sets (34 clips for Ped1 and 16 clips for Ped2), which contain only normal activities, for learning of normal models. Frame level ground truth, which indicates whether an anomaly is present in a frame, are provided for the testing sets (36 clips for Ped1 and 12 clips for Ped2). For a subset of the testing sets, pixel level ground truth masks are provided.

Due to the unsupervised nature of our method, we do not use the training sets but directly detect abnormal activities in the two testing sets. We only use several normal frames to partially calibrate the perspective distortion of the scene in Ped1 due to the use of optical flow as the low-level representation. For the first round of scan, the window size is fixed to 15×10 pixels in space and 3 frames in time. We compare our method with sparse reconstruction [5], mixture of dynamic textures (MDT) [15], social force (SF) [17], mixture



(a) (b) (c)

Figure 3. Comparison of anomaly detection results on UCSD dataset. (a) ROCs for Ped1 dataset using frame-level ground truth. (b) ROCs for Ped2 dataset using frame-level ground truth. (c) ROCs for a subset of Ped1 dataset using pixel-level ground truth. A frame is considered detected correctly only if at least 40% of the truly anomalous pixels are detected.

(a) Ped1

(b) Ped2

Figure 4. Examples of abnormal activities detected in UCSD anomaly detection dataset.

of optical flow (MPPCA) [12] and the optical flow monitoring method [1]. Note that except ours, all other methods have used the training sets to learn the normalcy models.

We show the ROC curves in Fig. 3. We can see that without any training, our scan statistic method has achieved remarkable results. On Ped1, although its performance is a little inferior to the sparse reconstruction method when using frame level ground truth, its localization accuracy is much higher than it as shown in Fig. 3(c). On Ped2 we outperform all other methods. After computing the optical flow and obtaining the cumulative flow word histograms for the videos, the scan statistic method takes about 0.2 second per frame on a standard Pentium machine with 3GHz CPU and 4GB RAM. Other states-of-the-art methods are much slower, e.g. the MDT method takes 25 secs per frame[15] during detection. Some examples of detections are shown in Fig. 4.

5.3. Subway Surveillance Dataset

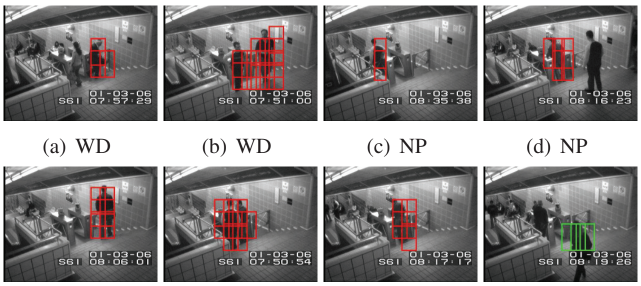
The subway surveillance videos are provided by Adam et al. [1]. There are two videos in the dataset. One video monitors the entrance gate, which is 1 hour 36 minutes long with 144,249 frames in total. The other watches the exit gate and is 43 minutes long with 64,900 frames. We follow the same definition of abnormal activities used in [12]. Specifically, the following abnormal activities are defined: (a) Wrong direction (WD): people exit through the entrance gate or enter through the exit gate; (b) No payment (NP): people enter the entrance gate without payment; (c) Loitering (LT): people loiter at the station; (d) Irregular interactions between persons (II); (e) Misc.: e.g. a person suddenly stops walking, or runs fast. The entrance gate video includes all these abnormal events and the exit gate video only includes the wrong direction, loitering and misc. events.

In this experiment, we use histogram of optical flow (HoF) to represent the local optical flow patch. The HoF features are then quantized to flow words. The spatial size of the window is fixed to 40×90 pixels and the length is set to 40 frames during the first round of scan. We then compare the detected merged regions with the ground truth. The quantitative comparison of the result with other methods is shown in Table 2. We can see that our algorithm achieved similar performance to other state-of-the-art methods. Unlike our method, both of the other two methods use video clips containing normal activities in the first few minutes of the video to train their models. Our method starts detection without training so that we do not need to know whether the gate is for entrance or exit beforehand. This is mined directly from the testing videos.

Our method missed half of the no payment events. This may due to the fact that the gate is located far from the camera and people turn their back towards the camera during paying. Additionally, some no payment actions are too subtle to be recognized. We show some examples of the detected abnormal activities in Fig. 5 and Fig. 6. The detected regions are composed of several overlapping sub-windows of fixed size. The composed windows show good coverage of the abnormal events.

Method	Wrong direction	No pay	Loitering	Irregular inter.	Misc.	Total	FA
Ground truth	26/9	13/-	14/3	4/-	9/7	66/19	0/0
ST-MRF [12]	24/9	8/-	13/3	4/-	8/7	57/19	6/3
Sparse coding [22]	25/9	9/-	14/3	4/-	8/7	60/19	5/2
Scan statistic	26/9	6/-	14/3	4/-	8/7	58/19	6/2

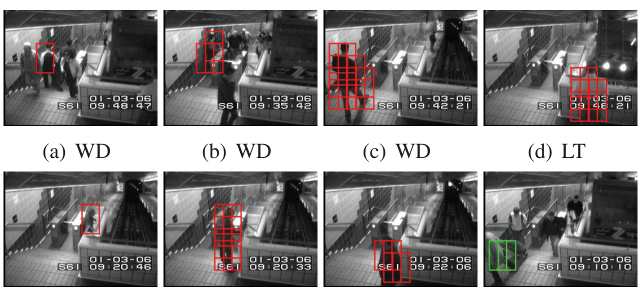
Table 2. Comparison of abnormal activity detection results on subway surveillance data. FA stands for false alarm. The first number in the slash (/) denotes the entrance gate result. The second is for the exit gate result.



(a) WD (b) WD (c) NP (d) NP

(e) LT (f) II (g) MISC (h) FA

Figure 5. Example of abnormal activities detected in subway entrance video. WD: wrong direction; NP: no payment; LT: loitering; II: irregular interactions; MISC: misc.; FA: false alarm. False alarms are marked with green windows.



(a) WD (b) WD (c) WD (d) LT

(e) MISC (f) MISC (g) MISC (h) FA

Figure 6. Example of abnormal activities detected in subway exit video. WD: wrong direction; LT: loitering; MISC: misc.; FA: false alarm. False alarms are marked with green windows.

5.4. Web Videos

The above experiments have demonstrated the effectiveness of our method on benchmark datasets. In this section, we apply our method to videos downloaded from the web which are taken in less controlled environments. In many web videos, the activities of the whole crowd keep changing and are unpredictable. In this case, the approaches that require normal examples for model training before actual detection become inapplicable. Our scan statistic method can still handle these videos. We first divide the videos into small clips. Then we detect abnormal activities in each clip independently.

Sample frames of the detected abnormal activities for three videos are shown in Fig. 7. In the first video, a man walks in the opposite direction to that of all other people. We correctly detect this man throughout the whole video. When the man is not in the scene, no abnormality occurs, so no alarm is raised. The second video is from a flash mob activity in which a group of people assemble and dance on the street. During the dancing, a man in a blue T-shirt suddenly steps out, dances in front of others and then goes back. We detect this man in most of the frames during his abnormal movement. A false alarm happens when his back is turned to the camera so the movement of his arms and legs are occluded. The third video displays crowd panic. Most people run in the scene. However, several people remain standing, then fall down. Our algorithm successfully detects these people in the video.

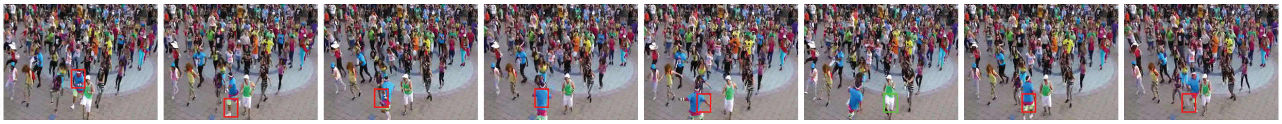
These results illustrate that the proposed method works well for real world videos in which crowd activities are complex. By using the criteria that abnormality happens when the behavior is different from the majority of others, we successfully detect abnormal activities in crowded scenes without any training.

6. Conclusion

In this work, we propose to use a semiparametric scan statistic method for abnormal activity detection in crowded scenes. Unlike many previous methods which need normal activity examples available beforehand for model training, our method is fully unsupervised and therefore more applicable in practice. A likelihood ratio test statistic is used as abnormality measurement for observations inside a scanning window. By using a semiparametric method to model the observations, it can be directly applied to analyze a broad variety of data. We have also proposed a fast scanning algorithm to improve the efficiency of the algorithm. Experimental results on four different datasets demonstrate the effectiveness of the proposed algorithm.

Scan statistics have become a popular method for cluster detection in many fields. However, their application to vision problems has not been exploited. In this work, we have illustrated its utility for abnormal crowd activity detection. It would also be interesting to use scan statistics to solve

(a) Crowd walk with a man walking in opposite direction.



(b) Crowd dance with a man stepping out.

(c) Crowd panic with several people falling down.

Figure 7. Abnormal activity detection result on three videos from the web. Red boxes on video frames represent detected abnormal activities. False alarms are marked with green boxes. Full videos are shown in the supplementary material.

other computer vision problems, such as abnormal object detection and texture analysis.

References

[1] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *PAMI*, 30:555–560, 2008.

[2] M. Bertini, A. D. Bimbo, and L. Seidenari. Multi-scale and real-time non-parametric approach for anomaly detection and localization. *CVIU*, 116:320–329, 2012.

[3] O. Boiman and M. Irani. Detecting irregularities in images and in video. In *ICCV*, 2005.

[4] M. D. Breitenstein, H. Grabner, and L. V. Gool. Hunting nessie – real-time abnormality detection from webcams. In *IEEE Int. Workshop on Visual Surveillance*, 2009.

[5] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR*, 2011.

[6] X. Cui, Q. Liu, M. Gao, and D. Metaxas. Abnormal detection using interaction energy potentials. In *CVPR*, 2011.

[7] K. Fokianos, J. Qin, B. Kedem, and D. A. Short. A semi-parametric approach to the one-way layout. *Technometrics*, 43(1):56–65, 2001.

[8] A. Hendel, D. Weinshall, and S. Peleg. Identifying surprising events in videos using bayesian topic models. In *ACCV*, 2010.

[9] http://mha.cs.umn.edu/Movies/Crowd Activity-All.avi.

[10] Y. Ito, K. Kitani, J. A. D. Bagnell, and M. Hebert. Detecting interesting events using unsupervised density ratio estimation. In *ECCV Workshop*, 2012.

[11] B. Kedem and S. Wen. Semi-parametric cluster detection. *Journal of statistical theory and practice*, 1(1):49–72, 2007.

[12] J. Kim and K. Grauman. Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates. In *CVPR*, 2009.

[13] M. Kulldorff. A spatial scan statistic. *Communications in Statistics – Theory and Methods*, 26(6):1481–1496, 1997.

[14] C. Liu. Beyond pixels: Exploring new representations and applications for motion analysis. *Doctoral Thesis.*, 2009.

[15] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, 2010.

[16] R. Mehran, B. E. Moore, and M. Shah. A streakline representation of flow in crowded scenes. In *ECCV*, 2010.

[17] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009.

[18] J. Qin and B. Zhang. A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3):609–618, 1997.

[19] X. Wang, X. Ma, and E. Grimson. Unsupervised activity perception by hierarchical bayesian models. In *CVPR*, 2007.

[20] S. Wu, B. E. Moore, and M. Shal. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *CVPR*, 2010.

[21] T. Xiang and S. Gong. Video behavior profiling for anomaly detection. *PAMI*, 30:893–908, 2008.

[22] B. Zhao, L. Fei-Fei, and E. P.Xing. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR*, 2011.

[23] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, 2004.