Video Anomaly Detection Based on Local Statistical Aggregates *

Venkatesh Saligrama Zhu Chen
Department of Electrical and Computer Engineering
Boston University, Boston, MA 02215
{srv,zchen}@bu.edu

Abstract

Anomalies in many video surveillance applications have local spatio-temporal signatures, namely, they occur over a small time window or a small spatial region. The distinguishing feature of these scenarios is that outside this spatio-temporal anomalous region, activities appear normal. We develop a probabilistic framework to account for such local spatio-temporal anomalies. We show that our framework admits elegant characterization of optimal decision rules.

A key insight of the paper is that if anomalies are local optimal decision rules are local even when the nominal behavior exhibits global spatial and temporal statistical dependencies. This insight helps collapse the large ambient data dimension for detecting local anomalies. Consequently, consistent data-driven local empirical rules with provable performance can be derived with limited training data. Our empirical rules are based on scores functions derived from local nearest neighbor distances. These rules aggregate statistics across spatio-temporal locations & scales, and produce a single composite score for video segments. We demonstrate the efficacy of our scheme on several video surveillance datasets and compare with existing work.

1. Introduction

Video surveillance has been an area of significant interest in both academia and industry. Recently, anomaly detection for video surveillance has gained importance [2, 7, 15, 8, 5, 14, 9, 11, 3, 21, 13, 16]. Our focus is on problems, where we are given a set of nominal training videos samples. Based on these samples we need to determine whether or not a test video contains an anomaly. We consider anomalies in motion attributes. Such outliers can include (un)usual motion patterns of (un)usual objects in (un)usual locations. These encompass anomalies such as dropped baggage, illegal U-turns, and sudden movements.

We focus on anomalies that have local spatio-temporal signatures. By locality we mean that the spatio-temporal region surrounding the anomalous region appears to follow the nominal activity and carries little information about the anomaly itself. For instance, the appearance of a bicyclist as shown in Fig. 1 illustrates spatio-temporal locality. As is seen outside a small window in time or in space the optical flow magnitudes look remarkably similar to nominal activity. We also consider other cases where locality is only temporal. These include cases such as sudden crowd movement [1] or illegal U-turns [5].

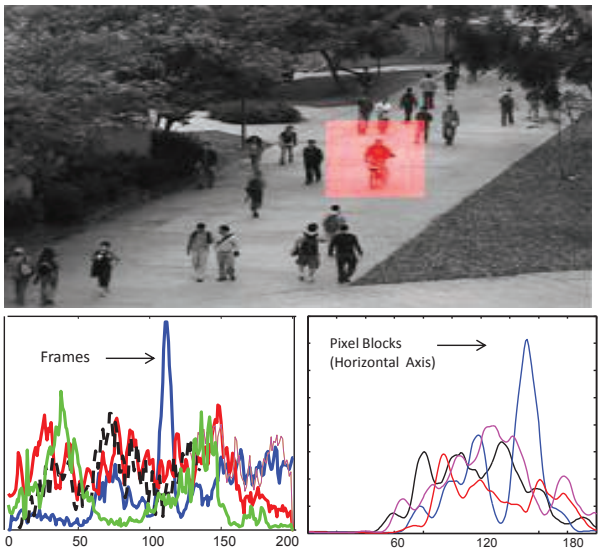


Figure 1. Illustration of local anomaly. Top: Illustrates frame of a video segment [15] with anomaly (bicycle). Bottom Panel (Left): Optical flow magnitude averaged over the red block vs. frame number for nominal and anomalous video segments. (Right): Optical flow magnitude averaged over different blocks along horizontal pixel blocks for different nominal and anomalous video. The magnitude outside "anomalous region" looks similar to nominal in both space and time.

We exploit these ideas by building on recent statisti-

*Research supported by ONR grant N000141010477, NGA grant HM1582-09-1-0037, NSF grant CCF-0905541, and DHS grant 2008-ST-061- ED0001

cal non-parametric notion of locality [18] and derive data-driven rules for video anomaly detection with predictable performance and statistical guarantees. Our approach is related to a number of other non-parametric data-driven approaches such as [19, 23] with key differences. Existing statistical approaches do not account for local anomalies, i.e., anomalies that are localized to a small time interval and/or spatial region. Our statistical locality notion leads to an elegant characterization of anomaly detection and suggests novel empirical rules. A fundamental insight gained from theoretical results is that the optimal decision rules for local anomalies are local irrespective of the global statistical dependencies exhibited in the nominal behavior. This insight helps collapse the large ambient data dimension for detecting local anomalies. Consequently, consistent data-driven local empirical rules with provable performance can be derived with limited training data. Our local empirical rules fuse local statistics and produce a composite score for a video segment. Anomalies are declared by ranking composite scores for video segments.

The paper is organized as follows. In Sec. 2 we present overview of our work and describe related work in video anomaly detection. In Sec. 3 we present our locality model structure to account for local spatio-temporal anomalies. Sec. 4 describes the Neyman-Pearson characterization and derives composite scoring schemes that lead to guarantees on false alarm control. Sec. 5 presents empirical rules that approximate the theoretical composite scores. Proofs of all statements appear in the supplementary section. Sec. 6 presents simulations on benchmark video datasets as well as comparisons to existing work.

2. Overview and Related Work

Our anomaly detection algorithm is described in Fig. 2. Our setup extracts local low-level motion descriptors and resembles other common approaches. Adam et al.[2] use histograms of optical flows at specific "local monitors" to derive decision rules for anomaly detection at those locations. Itti and Baldi consider low-level feature descriptors at every location [10] and use possion statistics for modeling nominal activity.

We propose a joint probability distribution of the low-level motion descriptors under nominal as well as anomalous distributions. Such joint distributions have also been considered extensively. Kim et al. [13] also extract local optical flow and enforce consistency across locations through Markov Random Field models. Benezeth et al. [5] use binary background subtraction to extract motion labels and then model these local features using a 3D Markov Random Field (MRF). Kratz et al. [14] extract spatio-temporal gradient to fit Gaussian model, and then use HMM to detect abnormal events. Mahadevan et al. [15] model the normal crowd behavior by mixtures of dynamic textures.

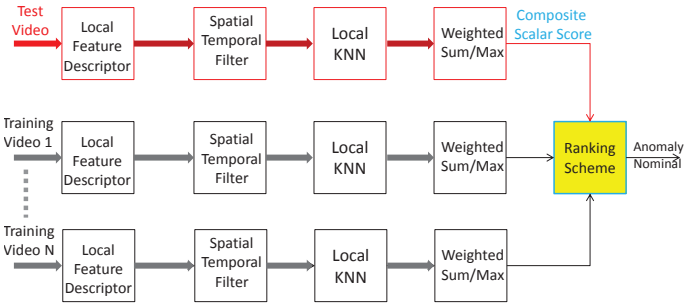


Figure 2. Overview of Anomaly Detection Algorithm. Motion descriptors are first extracted and quantized into small blocks. Spatio-Temporal filters at different scales are applied to obtain smooth estimates at each spatio-temporal location for each feature descriptor. Local KNN distance for each location is computed for training and test video. These local KNN distances are aggregated to produce a composite score for the test and training video. The composite scores are ranked to determine anomalies.

We introduce novel structural assumptions on the joint distributions to account for spatial and temporal locality of anomalies. Our locality assumption leads us to consider statistics on local 3D brick patches (space-time blocks) across different overlapping locations. These statistics are obtained through spatio-temporal filters as shown in Fig. 2. Our 3D modeling superficially resembles Boiman and Irani [7] but is different. They consider ensembles of 3D bricks and derive Gaussian models for matching test ensembles at a specific location with corresponding ensembles in a database. However, our goal is statistical and does not attempt to match 3D bricks at a location. Rather (see Fig. 2) we first compute location specific K-nearest neighbor (NN) distance for each 3D brick. We then normalize and compute a composite score by aggregating weighted K-NN distances from all the locations. This composite score is ranked against other such composite scores associated with training video segments. We then declare low scores as anomalies. It turns out that fusing local 3D brick statistics in this manner has theoretical significance. The empirical composite scoring and ranking scheme asymptotically converges to the optimal decision rule for maximizing detection power subject to false alarm constraints.

Our work is also related to Cong et. al. [8] who consider dictionary learning methods. There 3D patches with specific temporal and spatial scale are chosen to match each scenario. A dictionary of representative patterns are learnt based on training video. Anomalies are declared if the test sample cannot be represented using a sparse set of dictionary patterns. It is worth mentioning that we could incorporate their ideas into our scheme. Sparse decomposition for each spatio-temporal scale can be viewed as a feature vector that feeds into our local KNN block (see Fig. 2).

3. Spatio-Temporal Locality Model

We first describe an abstract problem and specialize it to video setting in Sec. 3.1. Consider a collection of random vectors, $x = (x_v)_{v \in V}$, indexed on a graph $G = (V, E)$. The set V is endowed with the usual graph metric $d(u, v)$ defined for any two nodes v and u.

We assume that baseline data $x = (x_v)_{v \in V}$ is drawn from the null hypothesis H_0:

$$H_0 : x \sim f_0(x) \tag{1}$$

We describe the anomalous distribution as a mixture of location and scale specific anomalous likelihood models. For simplicity of exposition we only consider location specific mixtures at a fixed scale s. Nevertheless, the techniques developed here can be generalized to mixtures across scales[1]. To this end, let $f_v(x)$, P_v be the likelihood function and prior probability associated with location v at scale s. Then,

$$H_1 : x \sim \sum_{v \in V} P_v f_v(x) \tag{2}$$

We next introduce notation to describe our local model. Let $\omega_{v,s}$ be a ball of radius s around v:

$$\omega_v \triangleq \omega_{v,s} = \{u \mid d(u, v) \leq s\}$$

With abuse of notation ω_v will generally refer to a ball of a fixed radius s at node v. We also denote by $\omega_{v,\epsilon}$ as the set that includes all points within an ϵ radius of ω_v, i.e.,

$$\omega_{v,\epsilon} = \{u \in V \mid d(u, v) \leq \epsilon,\ v \in \omega_v\}$$

The marginal distribution of f_0, f_v on a subset $\omega \subset V$ is denoted as $f_0(x_\omega)$.

Definition 1. We say an anomaly is of *local structure* if the distributions f_0 and f_v satisfy the following Markovian and Mask assumptions.

(1) Markov Assumption: We say f_0 and f_v's satisfy the Markov assumption if the observation x forms a Markov random field. Specifically we assume that there is an ϵ-neighborhood such that x_v, $v \in \omega_v$ is conditionally independent of x_u, $u \notin \omega_{v,\epsilon}$ when conditioned on the annulus $\omega_{v,\epsilon} \cap \omega_v^c$.

(2) Mask Assumption: The marginal distribution of f_0 and f_v on ω_v^c is identical:

$$f_0(x_{\omega_v^c}) = f_v(x_{\omega_v^c})$$

[1] $H_1 : x \sim \sum_s \sum_{v \in V} P_{v,s} f_{v,s}(x)$ where $P_{v,s}, f_{v,s}(x)$ are likelihood function and prior probability at location v and scale s.

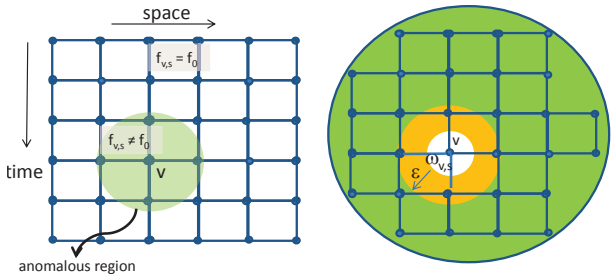


Figure 3. Illustration of Markov and Mask Properties. Markov implies random variables in region $\omega_{v,s}$ are independent of random variables in ω_v^c when conditioned on the annulus. Mask assumption means that the anomalous density and nominal density are identical outside ω_v.

3.1. Video Locality Model and Feature Descriptors

A video snippet x is typically a short segment of video. Training data can consist of several snippets, $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$. For theoretical purposes we assume that the different snippets are independent of each other. These snippets can be obtained by partitioning a longer video into short non-overlapping segments.

For a video snippet, x, we associate a graph $G = (V \times T, E)$. The set V is associated with spatial locations and the set T is associated with temporal locations in the video snippet. Each location, $v \in V$ and time $t \in T$ is associated with a feature descriptor $x_{v,t}$. While it is theoretically possible to consider all pixel locations and temporal instants, we quantize into $10 \times 10 \times 5$ non-overlapping blocks. We call these blocks as atoms and we associate average values of features for each atom. Two atoms are connected if they are either temporal or spatial neighbors. The rest of development with regards to Mask and Markov assumptions follow as in the previous section (also see Fig. 4).

Feature Descriptors: We now describe local features that are associated with each node (atom) of our graph. During feature extraction we compute a feature value for each pixel. Then, the pixel-level features are condensed into a multi-dimensional vector for each atom by averaging each feature component over all the pixels within the atom. We use the following local features:

(1) Persistence: Activity is detected using a basic background subtraction method (as for instance in [5]). The initial background is estimated using median of several hundred frames. Then, the background is updated using the running average method. We flag each pixel as part of the background or foreground. Persistence, for an atom, is the percentage of foreground pixels in the atom.

(2) Direction: Motion vectors are extracted using Horn and Schunck's optical flow method [6]. Motion is quantized into 8 directions and an extra "idle" bin is used for flow vectors with low magnitude. The feature for each atom is a 9-bin *un-normalized* motion histogram. The value for each

bin corresponds to the number of pixels moving in the direction associated with the bin.
(3) Motion Magnitude: Magnitude of motion vectors for each bin (except the idle bin) is computed and averaged over all the pixels in the atom.

We thus have an 11-dimensional descriptor for each atom. While our setup is sufficiently general and admits other descriptors we use only these in this paper.

4. Neyman-Pearson Characterization

We drop the explicit notation that indexes space and time described in Sec. 3.1 for notational convenience. Thus we are given a graph $G = (V, E)$ and associated features x_v for $v \in V$. An anomaly detector is a decision rule, π, that maps observations $x = (x)_{v \in V}$ to $\{0, 1\}$ with zero denoting no anomaly and one denoting an anomaly. Let $\Omega_\pi = \{x \mid \pi(x) = 1\}$. The optimal anomaly detector π minimizes the "Bayesian" Neyman-Pearson objective function.

$$\text{Bayesian:} \quad \max_{\pi} \int_{\Omega_\pi} \sum_{v \in V} P_v f_v(x) dx \tag{3}$$

$$\text{subject to}$$

$$P_F \triangleq \int_{\Omega_\pi} f_0(x) dx \leq \alpha$$

The optimal decision rule can be characterized as

$$\sum_v P_v \mathcal{L}_v \underset{nominal}{\overset{anomaly}{\gtrless}} \xi \tag{4}$$

where the likelihood ratio function $\mathcal{L}_v$ is defined as $\mathcal{L}_v = f_v(x)/f_0(x)$ and ξ is chosen such that the false alarm probability is smaller than α. Lemma 1 (see Supplementary Section for the proof) shows that the likelihood ratio function $\mathcal{L}_v$ simplifies under our assumptions of Definition 1.

Lemma 1. *Let ω_v be a ball around v and $\omega_{v,\epsilon}$ be the ϵ-neighborhood set such that the Markovian assumption of Definition 1 is satisfied. Then we have,*

$$\mathcal{L}_v(x) = \frac{f_v\left(x_{\omega_{v,\epsilon}}\right)}{f_0\left(x_{\omega_{v,\epsilon}}\right)} \tag{5}$$

Several issues arises in applying this decision rule. Both P_v and the likelihood model f_v are unknown and we only have nominal training data. A uniform prior ($P_v = 1/|V|$) or a worst case prior are options for dealing with unknown P_v. The worst-case prior turns out to be uniform under under symmetrizing location invariance assumptions. The issue of unknown f_v is an important aspect in anomaly detection and we follow the conventional practice and assume a uniform distribution over the support of $f_0(\cdot)$. Now $\mathcal{L}_v(x)$ is location dependent since support of f_0 varies with location. To account for this situation, we suppose that at location v, the collection of features $x_{\omega_{v,\epsilon}}$ corresponding to the spatial ball, $\omega_{v,\epsilon}$, lies in a set of diameter λ_v in the feature space. Note from before that the spatial ball $\omega_{v,\epsilon}$ has a spatial diameter $s + \epsilon$. With this notation, Eq. 5 reduces to:

$$\mathcal{L}_v(x) = \frac{\lambda_v^{-(s+\epsilon)}}{f_0\left(x_{\omega_{v,\epsilon}}\right)} \tag{6}$$

4.1. Composite Scores with Guarantees

While Equation 4 characterizes the optimal decision rule, it is unclear how to choose a threshold to ensure false alarm control. To this end we let $G(x)$ be a real-valued statistic of the raw data. Consider the score function:

$$R(\eta) = \mathbb{P}_{x \sim f_0}\left(x : G(x) \geq G(\eta)\right) \tag{7}$$

It is easy to show that this score function is distributed uniformly for a large class of statistics $G(x)$. This includes:
(1) NP detector: $G_{SUM}(x) = \sum_v \mathcal{L}_v(x)$.
(2) GLRT [12]: $G_{MAX}(x) = \max_v \mathcal{L}_v(x)$.
(3) Entropy: $G_{ENT}(x) = -\sum_v \log(\mathcal{L}_v(x))$.

Lemma 2. *Suppose statistics $G(x)$ has the nestedness property, that is, for any $t_1 > t_2$ we have $\{x : G(x) > t_1\} \subset \{x : G(x) > t_2\}$. Then $R(\eta)$ is uniformly distributed in $[0, 1]$ when $\eta \sim f_0$.*

This lemma implies that we can control false alarms via thresholding the statistic $R(\eta)$.

Theorem 3. *If G satisfies the nestedness property, by setting the detection rule as $R(\eta) \leq \alpha$, we control the FA at level α. Furthermore, if $R(\eta)$ is computed with $G_{SUM}(x) = \sum_v \mathcal{L}_v(x)$, then it is optimal solution to Equation 3 for the uniform prior.*

5. Empirical Composite Scores

The goal in this section is to empirically approximate $R(\cdot)$ given training data $(x^{(1)}, \cdots, x^{(n)})$, a test point η and a statistic $G_n(\cdot)$. Consider the empirical score function:

$$R_n(\eta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{G_n(x^{(i)}) \geq G_n(\eta)\}} \tag{8}$$

where G_n is a finite sample approximation of G and $\mathbb{I}_{\{\cdot\}}$ is the indicator function. Here we propose local nearest neighbor based statistics and the reasons for this choice will be described shortly. We denote it as a local neighborhood based composite score (LCS). This is because $G_n(\cdot)$ as described in the previous section combines statistics over local neighborhoods of a data sample and the ranking function produces a composite score for an entire random field.

Definition 2. We define the d-statistic $d_{\omega_{v,\epsilon}}(\eta)$ for window $\omega_{v,\epsilon}$ at an arbitrary point η as the distance of $\eta_{\omega_{v,\epsilon}}$ to its k-th closest point in $\left(x^{(1)}_{\omega_{v,\epsilon}}, \cdots, x^{(n)}_{\omega_{v,\epsilon}}\right)$.

We generally choose Euclidean distance for computing the distances. In general, we can apply any distance metric customized to specific application. To approximate $G(x)$ for different cases we need to determine the support parameter λ_v. To this end we let $d^{(j)}_{\omega_{v,\epsilon}}$ as the ordered the distances of $d_{\omega_{v,\epsilon}}(x^{(j)})$ $(j = 1, 2, \cdots, n)$ in decreasing order and we approximate the support as an ξ percentile:

$$\lambda_v = d^{(\lfloor n\psi \rfloor)}_{\omega_{v,\epsilon}} \tag{9}$$

where $\lfloor n\psi \rfloor$ denotes the integer part of real number and can be tuned (in simulations we usually use the 95th percentile). Now $G_{SUM}(x) = \sum_v \mathcal{L}_v(x)$ can be approximated by SUM LCS, $G_{n,SUM}$:

$$G_{n,SUM} = \sum_v \left(\frac{d_{\omega_{v,\epsilon}}(\eta)}{\lambda_v}\right)^s \tag{10}$$

Similarly, we can take a max statistic to obtain MAX LCS:

$$G_{n,MAX} = \max_v \frac{d_{\omega_{v,\epsilon}}(\eta)}{\lambda_v} \tag{11}$$

Observe that when s is equal to dimension of x the two statistics max and sum coincide. Inverse of the nearest neighbor distance have concentrate around the density [17] and motivate our choice for using such distances. Our resulting LCS statistics Eq. 8 is identical to the K-nearest-neighbor ranking (KNN-Ranking) scheme of [23] for anomaly detection.

Practical Issues with $G_{SUM}(\cdot)$: Recall from Section 4, $G_{SUM}(x)$, appears to be optimal for uniform priors and minimax optimal under symmetrizing assumptions. However, it is difficult to reliably approximate $G_{SUM}(x)$ for several reasons. (1) Sum is no longer optimal if the prior is not uniform. (2) Errors can accumulate for the summation but max is relatively robust. (3) The additional s exponent term in the expression of SUM LCS (which compensates for the dimension) leads to sensitivity to parameters such as λ_v. (4) For large values of s, since max distance is a dominant term in $G_{SUM}(x)$ the theoretical difference between the two statistics maybe negligible. Therefore, we adopt MAX-LCS in this paper.

Theoretical Properties The theoretical properties for MAX-LCS and SUM-LCS are described in [18]. We provide some of the main results for MAX-LCS here. It turns out that under sufficient smoothness conditions on $f_0(\cdot)$, if $\eta \sim f_0$, then the score $R_{n,MAX}(\eta)$ converges to a uniform distribution on the unit interval.

$$R_{n,MAX}(\eta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{G_{n,MAX}(x^{(i)}) \geq G_{n,MAX}(\eta)\}} \xrightarrow{d} U[0,1]$$

Consequently, to control false alarms at level α asymptotically our decision rule is to:

$$R_{n,MAX}(\eta) \underset{anomaly}{\overset{nominal}{\gtrless}} \alpha$$

6. Experiments and Comparisons

To test the performance of our proposed algorithm, we apply it to several published datasets and compare our results with existing work. We used the UCSD dataset [20], the UMN dataset [1] of crowd anomalies, the Uturn dataset [5] and the Subway dataset [2].

6.1. Algorithm for Video Anomaly Detection

Recall we are given training video samples and a test video sample. To reduce real-time delay we breakup the test video sample into test video snippets, $\eta^{(1)}, \eta^{(2)}, \ldots, \eta^{(m)}$. Our task is to determine which of the test snippets contain an anomaly. For convenience, we partition training video in to snippets $(x^{(1)}, \cdots, x^{(n)})$ each of the same length as a test snippet $\eta \triangleq \eta^{(j)}$. Our algorithm consists of three steps:

(1) Local Scores: For any snippet y, which denotes either a test or training snippet, a local score at spatial location v, temporal instant, t, and at spatio-temporal scale s, is computed (see Algorithm 1). We choose a uniform spatio-temporal filter with support equal to s for simplicity in Algorithm 1.

Algorithm 1 Score for y at location (v, t), at scale s.

Input: $\{x^{(j)}\}$; KNN parameter: K, location:(v, t); Scale: s

Output: $d_{y,v,t}(s)$

1: Filter at scale s: $x^{(j)} \leftarrow Filter_s(x^{(j)})$
2: Distance Computation: $d_j \leftarrow \text{dist}(y_{v,t}, x^{(j)}_{v,\tau})$, $\forall\, j,\ \tau$
3: Compute $d_{(\ell)}$ the ℓth nearest neighbor distance by sorting d_j.
4: Average:$d_{v,t} \leftarrow \frac{1}{K}\sum_{\ell=K+1}^{2K} d_{(\ell)}$
5: Normalize $d_{y,v,t}(s) \leftarrow \frac{d_{v,t}(s)}{D_v}$; where $D_v = \max_t d_{v,t}$

(2) Snippet Score: Compute composite score for each snippet—test and training snippets—from local scores obtained in Algorithm 1:

$$d_y(s) = \max_{v,t} d_{y,v,t}(s) \tag{12}$$

(3) Anomaly Detection: Rank test snippet, η at scale s:

$$R_s(\eta) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_{\{d_{x^{(j)}}(s) > d_\eta(s)\}}$$

Note that our feature descriptors–magnitude, direction, persistency—have different dynamic ranges. In this paper

we ranked separately with respect to the different descriptors. Anomalies are declared if the score at scale s, $R_s(\eta)$, for any descriptor falls below the desired false alarm threshold, α. If an anomaly for a snippet is declared, the anomaly is localized by identifying the spatio-temporal locations, v, t in the snippet that achieve the maximum in Eq. 12.

Tuning Parameters: Our algorithm requires only two parameters, namely, K for KNN distance computation and scale s. It turns out that our results are generally robust to a wide range of K and is not an issue. In all our simulations we choose K to be about 50. Scale s can be dealt with in two possible ways: (**1**) Compute ranks over different scales and declare anomaly if the rank at some scale falls below the threshold. This procedure is conservative; Nevertheless, it controls false alarms at desired level asymptotically. (**2**) Use context to determine sensible temporal and spatial scales. This idea has been used before by Cong et. al. [8], who choose appropriate basis depending on the scenario. We choose small scales if small scale anomalies (abandoned or unusual objects) are important and choose larger scales for spatial anomalies such as U-turns or global change in behavior.

Computational Issues: KNN distance computation is our main bottleneck. It scales linearly with the number of 3D bricks. To overcome this drawback recent approaches for computing approximate nearest neighbors based on locality sensitive hashing(LSH) [4] can be used. While we do not present results based on LSH here, in our preliminary experiments we have noticed that it can drastically reduce the computation time (scaling as fourth root of the number of 3D bricks) with little loss in performance.

6.2. UCSD Ped1 dataset [20]

The UCSD Ped1 dataset contains 34 training clips of nominal patterns and 36 testing clips of various abnormal events, *e.g*. bicycles, skaters, carts, etc. Each clip has 200 frames (20 seconds), with a 158 $\times$ 238 resolution. The challenge in this dataset is that the scenes are extremely crowded. To apply our algorithm, first we calculated optical flow and aggregated optical flow into histogram and magnitude features. We divided the videos into overlapping spatio-temporal blocks of 30pixels$\times$20pixels$\times$5frames (the block size was chosen such that each block does not contain too many objects which may interfere with one another) and then we applied our algorithm on snippets consisting of 5 frames. We also experimented with larger snippets and noticed little performance degradation.

Some image results are shown in Figure 4. Our algorithm can detect different types of anomalies. In Figure 5, we compared ROC curves of our method with SRC proposed in [8] and MDT proposed in [15]. We also compared our method with Social force and MPPCA, etc. It is easy to see that our method outperforms all the other algorithms. In

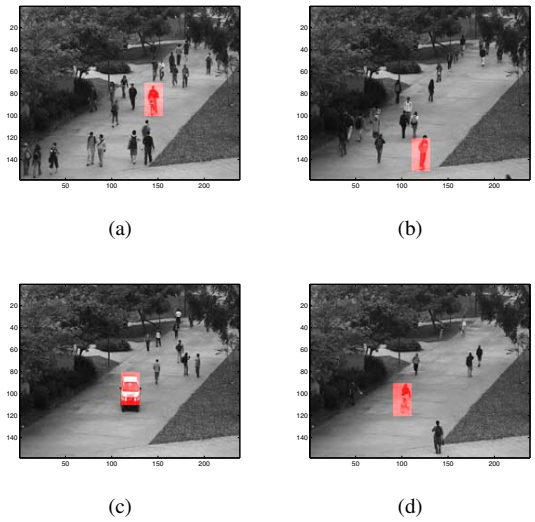


Figure 4. Abnormal event detections for UCSD Ped1 datasets. The objects such as cars, bicycles, skaters are all well detected.

Table 1, some evaluation results are presented: the Equal Error Rate (EER) (ours 16% $<$ 19% [8]), and Area Under Curve (AUC) (ours 92.7% $>$ 86% [8]). From these comparisons, we can conclude that our algorithm outperforms other state-of-the-art algorithms. One additional advantage of our algorithm is that while providing frame level results, we can also provide anomaly localization by back-tracing to the block with max statistics.

Method	EER	AUC
MPPCA [15]	40%	59%
SF [15]	31%	67.5%
MDT [15]	25%	81.8%
Sparse [8]	19%	86%
Ours	**16%**	**92.7%**

Table 1. Quantitative comparison of our algorithm with [8] and [15]. EER is equal error rate and AUC is the area under ROC.

6.3. Subway dataset [2]

The subway dataset is obtained from Adam *et al*. [2]. In our experiments, we used the "entrance gate" video which is 1 hour 36 minutes long with 144249 frames. For our experiments, we applied a 96pixels $\times$96pixels $\times$50frames block (2 seconds).

In Figure 6, a few detected abnormal frames are shown with abnormal blocks marked red. In Figure 7 we compare the frame level ROC curves with results in [8]. It is obvious that our algorithm outperforms SRC proposed in [8] significantly for the subway dataset.

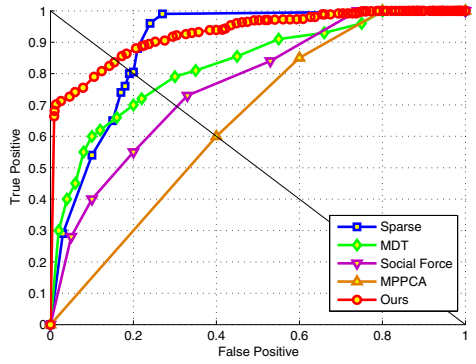
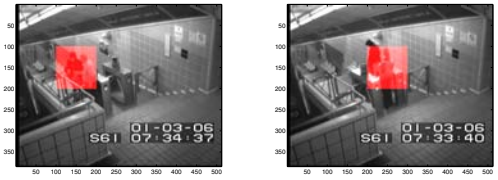


Figure 5. The detection results of UCSD Ped1 dataset.



(a) (b)

Figure 6. Abnormal event detections for Subway datasets.

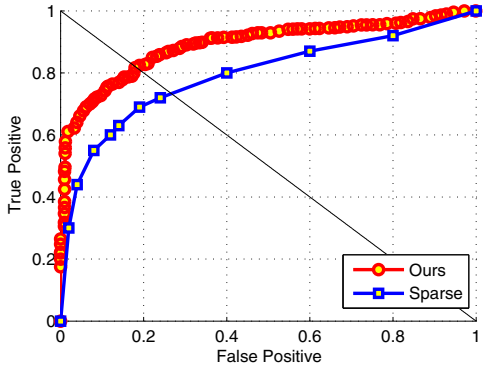
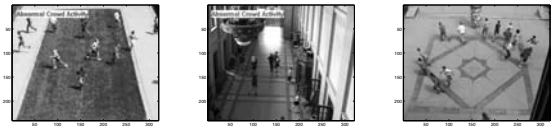


Figure 7. The detection results of Subway dataset.

6.4. UMN dataset [1]

The UMN dataset [1] consists of 3 different scenes of crowds of walking people who suddenly started running. Scene 1 contains 1450 frames, scene 2 contains 4415 frames and scene 3 contains 2145 frames all with a 320×240 resolution. We used a 80pixels$\times$80pixels$\times$45frames block and trained our algorithm using first 600 frames of each scene and use the others for testing.

In Figure 8, we demonstrate some detected abnormal frames using our proposed algorithm. Table 2 provides quantitative comparisons to other state-of-the-art algorithms. Our proposed algorithm is comparable to [22] and [8], and outperforms [16]. Note that our method is simpler and requires little parameter tuning in comparison to other methods. In Figure 9 we compare the ROC curves with several other algorithms.



(a) (b) (c)

Figure 8. Abnormal event detections for UMN datasets.

Method	AUC
Chaotic Invariants [22]	99%
Social Force [16]	96%
Optical Flow [16]	84%
Sparse [8]	97.5%
Ours	**98.5**%

Table 2. Quantitative comparison of our algorithm with [8], [16] and [22]. AUC is the area under ROC.

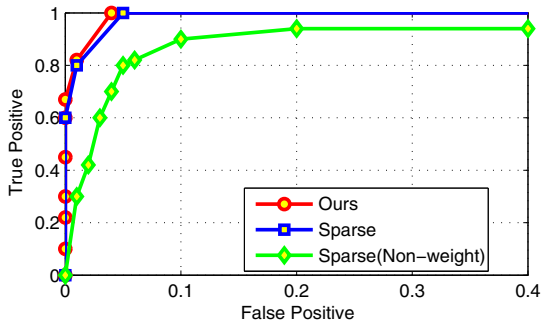
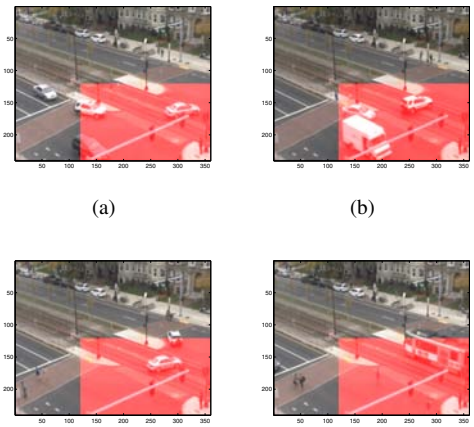


Figure 9. The detection results of UMN dataset.

6.5. Uturn dataset [5]

The Uturn dataset is made available to us by Benezeth *et al.* [5]. It is a video of a junction with cars driving in different directions, trams passing by and pedestrians walking about. The anomalous activities in this case are illegal U-turns and trams. The video contains 6057 frames. A block of 120pixels $\times$240pixels $\times$60 frames is adopted in our experiment. Anomalous frames are shown to illustrate the detected anomaly in Figure 10. The results are depicted in Figure 11. Also in the top panel of Fig. 11 illustrates how direction histograms behave as a time series. The first anomalous instance (marked as red in the truth) is the tram, where direction 1, 2, 7 and 8 has large intensity. The rest of the anomalous instances are Uturns, where the features

corresponding to the U-turn is spread out in all of the first 5 directions in the histogram. Both these types of anomalies are distinct from normal activities.



(a) (b) (c) (d)

Figure 10. Abnormal event detections for Uturn dataset. Depicts Uturn and Tram anomalies on the same spatio-temporal block on different frames.

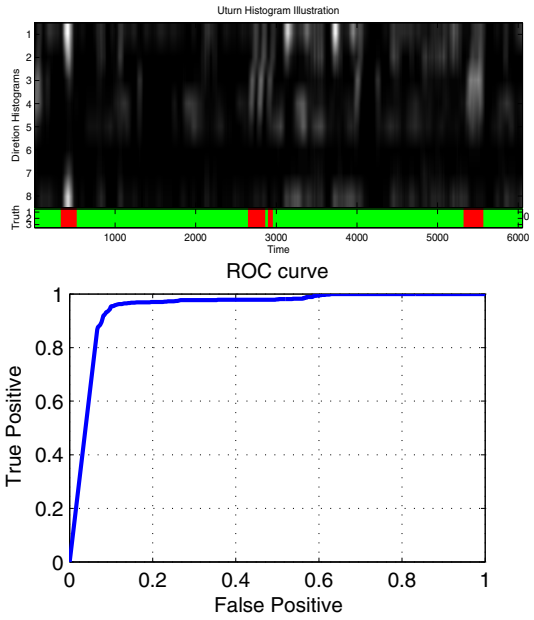


Figure 11. The detection results of Uturn dataset.

References

[1] Unusual crowd activity dataset. http://mha.cs.umn.edu/Movies/Crowd-Activity-All.avi. 1, 5, 7

[2] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real-time unusual event detection using multiple fixed-location monitors. *PAMI*, 30(3):555–560, 2008. 1, 2, 5, 6

[3] S. Ali and M. Shah. Floor fields for tracking in high density crowd scenes. *ECCV*, 2008. 1

[4] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM*, 51:117–122, January 2008. 6

[5] Y. Benezeth, P. Jodoin, V. Saligrama, and C. Rosenberger. Abnormal events detection based on spatio-temporal co-occurences. *CVPR*, 2009. 1, 2, 3, 5, 7

[6] B.Horn and B. Schunck. Determining optical flow. 17(1-3):185–203, 1981. 3

[7] O. Boiman and M. Irani. Detecting irregularities in images and in video. *ICCV*, 2005. 1, 2

[8] Y. Cong, J. Yuan, and J. Liu. Sparse reconstruction cost for abnormal event detection. *CVPR*, 2011. 1, 2, 6, 7

[9] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *PAMI*, 28(9):1450–1464, 2006. 1

[10] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR*, pages 631–637, 2005. 2

[11] F. Jiang, J. Yuan, S. Tsaftaris, and A. Katsaggelos. Anomalous video event detection using spatiotemporal context. *Computer Vision and Image Understanding*, 115(3):323–333, 2011. 1

[12] S. Kay. *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall, 1998. 4

[13] J. Kim and K. Grauman. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. *CVPR*, 2009. 1, 2

[14] L. Kratz and K. Nishino. Anomaly detection inextremely crowded scenes using spatio-temporal motion pattern models. *CVPR*, 2009. 1, 2

[15] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. *CVPR*, 2010. 1, 2, 6

[16] R. Mehran, A. Oyama, and M. Shah. Abnormal crowd behavior detection using social force model. *CVPR*, 2009. 1, 7

[17] J. Qian and V. Saligrama. Graph construction for learning on unbalanced data. http://arxiv.org/abs/1112.2319, 2011. 5

[18] V. Saligrama and M. Zhao. Local anomaly detection. In *AISTATS*, 2012. 2, 5

[19] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, July 2001. 2

[20] UCSD. Anomaly detection dataset, 2010. http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm. 5, 6

[21] X. Wang, X. Ma, and W. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *PAMI*, 31(3):539–555, 2009. 1

[22] S. Wu, B. Moore, and M. Shah. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. *CVPR*, 2010. 7

[23] M. Zhao and V. Saligrama. Anomaly detection with score functions based on nearest neighbor graphs. In *NIPS*, volume 22, 2009. 2, 5