Sparse Reconstruction Cost for Abnormal Event Detection

Yang Cong[1], Junsong Yuan[1], Ji Liu[2]
[1]School of EEE, Nanyang Technological University, Singapore
[2]University of Wisconsin-Madison, USA
congyang81@gmail.com, jsyuan@ntu.edu.sg, ji-liu@cs.wisc.edu

Abstract

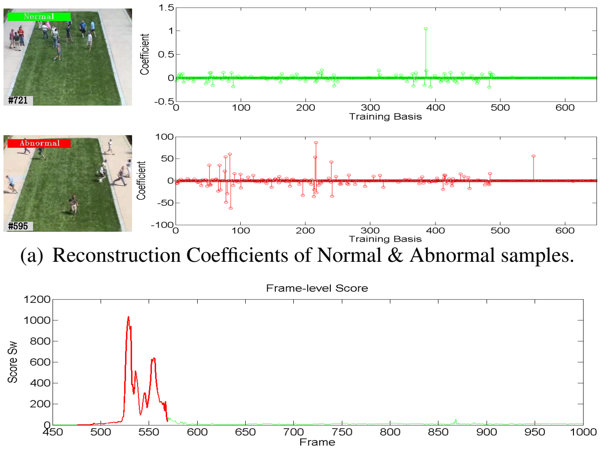
We propose to detect abnormal events via a sparse reconstruction over the normal bases. Given an over-complete normal basis set (e.g., an image sequence or a collection of local spatio-temporal patches), we introduce the sparse reconstruction cost (SRC) over the normal dictionary to measure the normalness of the testing sample. To condense the size of the dictionary, a novel dictionary selection method is designed with sparsity consistency constraint. By introducing the prior weight of each basis during sparse reconstruction, the proposed SRC is more robust compared to other outlier detection criteria. Our method provides a unified solution to detect both local abnormal events (LAE) and global abnormal events (GAE). We further extend it to support online abnormal event detection by updating the dictionary incrementally. Experiments on three benchmark datasets and the comparison to the state-of-the-art methods validate the advantages of our algorithm.

1. Introduction

The Oxford English Dictionary defines *abnormal* as:

> *deviating from the ordinary type, especially in a way that is undesirable or prejudicial; contrary to the normal rule or system; unusual, irregular, aberrant.*

According to the definition, the abnormal events can be identified as irregular events from normal ones. Thus, the task is to identify abnormal (negative) events given the normal (positive) training samples. To address this one-class learning problem, most conventional algorithms [2, 15, 14, 20] intend to detect testing sample with lower probability as anomaly by fitting a probability model over the training data. As a high-dimensional feature is essential to better represent the event and the required number of training data increases exponentially with the feature dimension, it is unrealistic to collect enough data for density estimation in practice. For example, for our global abnormal detection, there are only 400 training samples with dimension of 320. With such a limited training samples, it is difficult to even fit a Gaussian model. Sparse representation is suitable to represent high-dimensional samples, we thus propose to detect abnormal events via a sparse reconstruction from normal ones. Given an input test sample $\mathbf{y} \in \mathbb{R}^m$, we reconstruct it by a sparse linear combination of an over-complete normal (positive) basis set $\Phi = \mathbb{R}^{m \times D}$, where $m < D$. To quantify the normalness, we propose a novel *sparse reconstruction cost* (SRC) based on the weighted l_1 minimization. As shown in Fig.1, a normal event is likely to generate sparse reconstruction coefficients with a small reconstruction cost, while abnormal event is dissimilar to any of the normal basis, thus generates a dense representation with a large reconstruction cost.



(a) Reconstruction Coefficients of Normal & Abnormal samples.

(b) Frame-level SRC (S_w).

Figure 1. (a) Top left: the normal sample; top right: the sparse reconstruction coefficients; bottom left: the abnormal sample; bottom right: the dense reconstruction coefficients. (b) Frame-level Sparsity Reconstruction Cost (SRC): the red/green color corresponds to abnormal/normal frame, respectively. It shows that the SRC (S_w) of abnormal frame is greater than normal ones, and we can identify abnormal events accordingly.

Depending on the applications, we classify the abnormal events into two categories: the local abnormal event (LAE), where the local behavior is different from its spatio-temporal neighborhoods; or the global abnormal event (GAE), where the whole scene is abnormal, even though any individual local behavior can be normal. To handle both cases, the definition of training basis $\mathbf{y}$ can be quite flexible, such as image patch or spatio-temporal subvolume. It thus provides a general way of representing different types of abnormal events. Moreover, we propose a new dictionary selection method to reduce the size of the basis set Φ for an efficient reconstruction of $\mathbf{y}$. The weight of each basis is also learned to indicate its individual normalness, i.e., the occurrence frequency. These weights form a weight matrix $\mathbf{W}$ which serves as a prior term in the l_1 minimization.

We evaluate our method in three datasets and the comparison with the state-of-the-art methods validate the following advantages of our proposed methods:

- We take into account the prior of each basis as the weight for l_1 minimization and propose a criterion (SRC) to detect abnormal event, which outperforms the existing criterion, e.g., Sparsity Concentration Index in [25].
- Benefitting from our new dictionary selection model using sparsity consistency, our algorithm can generate a basis set of minimal size and discard redundant and noisy training samples, thus increases computational efficiency accordingly.
- By using different types of basis, we provide a unified solution to detect both local and global abnormal events in crowded scenes. Our method can also be extended to online event detection via an incremental self-update mechanism.

2. Related Work

Research in video surveillance has made great progresses in recent years, such as background model [22], object tracking [3], pedestrian detection [8], action recognition [27] and crowd counting [7]. Abnormal event detection, as a key application in video surveillance, has also provoked great interests. Depending on the specific application, the abnormal event detection can be classified into those in the crowded scenes and those in the un-crowded scenes. For the un-crowded scenario, binary features based on background model have been adopted, such as Normalization Cut clustering by Hua et al. [29] and 3D spatio-temporal foreground mask feature fusing Markov Random Field by Benezeth et al. [4]. There are also some trajectory-based approaches to locate objects by tracking or frame-difference, such as [10], [24], [21] and [13].

For the crowded scenes, according to the scale, the problem can be classified into LAE and GAE. Most of the state-of-the-art methods consider the spatio-temporal information. For LAE, most work extract motion or appearance features from local 2D patches or local 3D bricks, like histogram of optical flow, 3D gradient, etc; the co-occurrence matrices are often chosen to describe the context information. For example, Adam et al. [2] use histograms to measure the probability of optical flow in a local patch. Kratz et al. [15] extract spatio-temporal gradient to fit Gaussian model, and then use HMM to detect abnormal events. The saliency features are extracted and associated by graph model in [12]. Kim et al. [14] model local optical flow with MPPCA and enforce the consistency by Markov Random Field. In [23], a graph-based non-linear dimensionality reduction method is used for abnormality detection. Mahadevan et al.[18] model the normal crowd behavior by mixtures of dynamic textures.

For the GAE, Mehran et al. [20] present a new way to formulate the abnormal crowd behavior by adopting the social force model [9], and then use Latent Dirichlet Allocation (LDA) to detect abnormality. In [26], they define a chaotic invariant to describe the event. Another interesting work is about irregularities detection by Boiman and Irani [5, 6], in which they extract 3D bricks as the descriptor and use dynamic programming as inference algorithm to detect the anomaly. Since they search the current feature from all the features in the past, this approach is time-consuming.

3. Our Method

3.1. Overview

In this paper, we propose a general abnormal event detection framework using sparse representation for both LAE and GAE. The key part of our algorithm is the sparsity pursuit, which has been a hot topic in machine learning recently and includes cardinality sparsity [11], group sparsity [28], matrix or tensor rank sparsity [17]. Assisted by Fig.1-2, we will show the basic idea of our algorithm. In Fig.2(C), each point is a feature point in a high dimensional space; various features are chosen for LAE or GAE depending on the circumstances, which is concatenated by Multi-scale Histogram of Optical Flow (MHOF), as in Fig.2(B). Usually at the beginning, only several normal frames are given for initialization and features are extracted to generate the whole feature pool $\mathbf{B}$ (the light blue points), which contains redundant noisy points. Using sparsity consistency in Sec.3.5, an optimal subset $\mathbf{B}'$ with a small size is selected from $\mathbf{B}$ as training dictionary, e.g. dark blue points in Fig.2(C), where the radius of each blue point relates to its importance, i.e. its weight.

In Sec.3.4, we introduce how to test the new sample $\mathbf{y}$. Each testing sample $\mathbf{y}$ could be a sparse linear combination of the training dictionary by a weighted l_1 minimization. Whether $\mathbf{y}$ is normal or not is determined by the linear reconstruction cost S_w, as shown in Fig.1. Moreover, our system can also online self-update, as will be discussed in

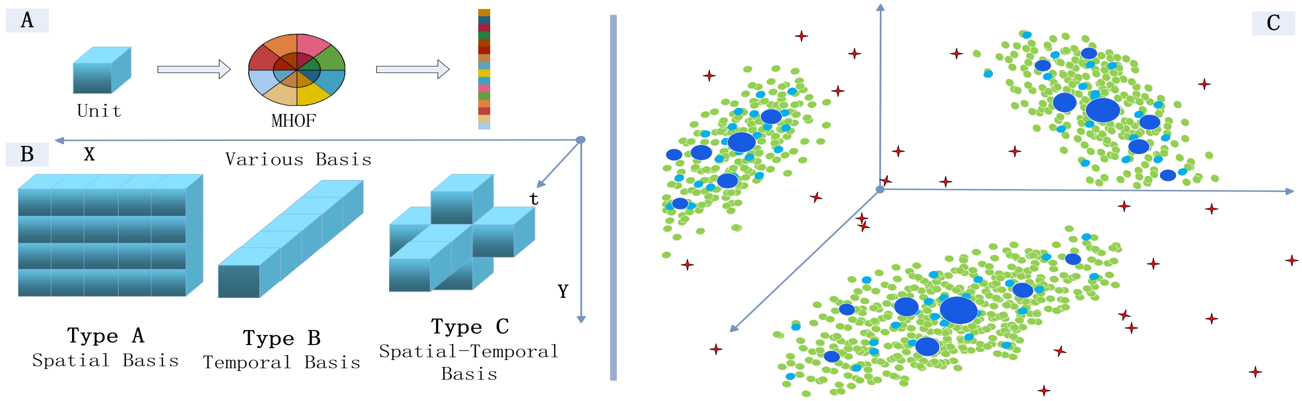


Figure 2. **(A)** The Multi-scale HOF is extracted from a basic unit (2D image patch or 3D brick) with 16 bins. **(B)** The flexible spatio-temporal basis for sparse representation, such as type A, B and C, concatenated by MHOF from basic units. **(C)** The illustration of our algorithm. The green or red point indicates the normal or abnormal testing sample, respectively. An optimal subset of representatives (dark blue point) are selected from redundant training features (light blue points) as basis to constitute the normal dictionary, where its radius indicates the weight. The larger the size, the more normal the representative. Then, the abnormal event detection is to measure the sparsity reconstruction cost (SRC) of a testing sample (green or red points) over the normal dictionary (dark blue points).

Sec.3.5. The Algorithm is shown in Alg.2.

3.2. Multi-scale HOF and Basis Definition

To construct the basis for sparse representation, we propose a new feature descriptor called Multi-scale Histogram of Optical Flow (MHOF). As shown in Fig.2(A), the MHOF has K=16 bins including two scales. The smaller scale uses the first 8 bins to denote 8 directions with motion magnitude $r < T_r$; the bigger scale uses the next 8 bins corresponding to $r \geq T_r$ (T_r is the magnitude threshold). Therefore, our MHOF not only describes the motion direction information as traditional HOF, but also preserves the more precise motion energy information. After estimating the motion field by optical flow [16], we partition the image into a few basic units, i.e. 2D image patches or spatio-temporal 3D bricks, then extract MHOF from each unit.

To handle different local abnormal events (LAE) and global abnormal events (GAE), we propose several bases with various spatio-temporal structures, whose representation by a normalized MHOF is illustrated in Fig.2(B). For GAE, we select the spatial basis covering the whole frame. For LAE, we extract the temporal or spatio-temporal basis that contains spatio-temporal contextual information, such as the 3D Markov Random Field [14]. And the spatial topology structure can take place the co-occurrance matrix. In general, our definition of the basis is very flexible and other alternatives are also acceptable.

3.3. Dictionary Selection

In this section, we address the problem of how to select the dictionary given an initial candidate feature pool as $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_k] \in \mathbb{R}^{m\times k}$, where each column vector $\mathbf{b}_i \in \mathbb{R}^m$ denotes a normal feature. Our goal is to find an optimal subset to form the dictionary $\mathbf{B}' = [\mathbf{b}_{i_1}, \mathbf{b}_{i_2}, \cdots, \mathbf{b}_{i_n}] \in \mathbb{R}^{m\times n}$ where $i_1, i_2, \cdots, i_n \in \{1, 2, \cdots, k\}$, such that the set $\mathbf{B}$ can be well reconstructed by $\mathbf{B}'$ and the size of $\mathbf{B}'$ is as small as possible. A simple idea is to pick up candidates randomly or uniformly to build the dictionary. Apparently, this cannot make full use of all candidates in $\mathbf{B}$. Also it is risky to miss important candidates or include the noisy ones, which will affect the reconstruction. To avoid this, we present a principled method to select the dictionary. Our idea is that we should select an optimal subset of $\mathbf{B}$ as the dictionary, such that the rest of the candidates can be well reconstructed from it. More formally, we formulate the problem as follows:

$$\min_{\mathbf{X}} : \frac{1}{2}\|\mathbf{B} - \mathbf{B}\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_1, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{k\times k}$; the Frobenius norm $\|\mathbf{X}\|_F$ is defined as $\|\mathbf{X}\|_F := (\sum_{i,j}\mathbf{X}_{ij}^2)^{\frac{1}{2}}$; and the l_1 norm is defined as $\|\mathbf{X}\|_1 := \sum_{i,j}|\mathbf{X}_{ij}|$. However, this tends to generate a solution of $\mathbf{X}$ close to $\mathbf{I}$, which leads the first term of Eq. 1 to zero and is also very sparse. Thus, we need to require the consistency of the sparsity on the solution, i.e., the solution needs to contain some "0" rows, which means that the corresponding features in $\mathbf{B}$ are not selected to reconstruct any data samples. We thus change the l_1 norm constraint in Eq. 1 into the $l_{2,1}$ norm, defined as $\|\mathbf{X}\|_{2,1} := \sum_{i=1}^{k}\|\mathbf{X}_{i.}\|_2$, where $\mathbf{X}_{i.}$ denotes the i^{th} row of $\mathbf{X}$. The problem is now formulated as:

$$\min_{\mathbf{X}} : \frac{1}{2}\|\mathbf{B} - \mathbf{B}\mathbf{X}\|_F^2 + \lambda\|\mathbf{X}\|_{2,1}. \tag{2}$$

The dictionary $\mathbf{B}'$ is constituted by selecting basis with $\|\mathbf{X}_{i.}\|_2 \neq 0$. The $l_{2,1}$ norm is indeed a general version of the l_1 norm since if $\mathbf{X}$ is a vector, then $\|\mathbf{X}\|_{2,1} = \|\mathbf{X}\|_1$. In addition, $\|\mathbf{X}\|_{2,1}$ is equivalent to $\|\mathbf{x}\|_1$ by constructing a new vector $\mathbf{x} \in \mathbb{R}^k$ with $\mathbf{x}_i = \|\mathbf{X}_{i.}\|_2$. From this angle, it is not hard to understand that Eq. 1 leads to a sparse solution for $\mathbf{X}$, i.e., $\mathbf{X}$ is sparse in terms of rows.

Next we show how to solve this optimization problem in Eq. 2, which is a convex but nonsmooth optimization problem. Since $\|\mathbf{X}\|_{2,1}$ is nonsmooth, although the general optimization algorithm (the subgradient descent algorithm) can solve it, the convergence rate is quite slow. Recently, Nesterov [19] proposed an algorithm to efficiently solve a type of convex (but nonsmooth) optimization problem and guarantee a convergence rate of $O(1/K^2)$ (K is the iteration number), which is much faster than the subgradient decent algorithm of $O(1/\sqrt{K})$. We thus follow the fundamental framework of Nesterov's method in [19] to solve this problem in Eq. 2. Consider an objective function $f_0(x)+g(x)$ where $f_0(x)$ is convex and smooth and $g(x)$ is convex but nonsmooth. The key technique of Nesterov's method is to use $p_{Z,L}(x) := f_0(Z) + \langle \nabla f_0(Z), x-Z\rangle + \frac{L}{2}\|x-Z\|_F^2 + g(Z)$ to approximate the original function $f(x)$ at the point Z. At each iteration, we need to solve $\arg\min_{x} : p_{Z,L}(x)$.

In our case, we define $f_0(\mathbf{X}) = \frac{1}{2}\|\mathbf{B}-\mathbf{B}\mathbf{X}\|_F^2$, $g(\mathbf{X}) = \lambda\|\mathbf{X}\|_{2,1}$. So we have

$$p_{\mathbf{Z},L}(\mathbf{X}) = f_0(\mathbf{Z}) + \langle \nabla f_0(\mathbf{Z}), \mathbf{X}-\mathbf{Z}\rangle + \frac{L}{2}\|\mathbf{X}-\mathbf{Z}\|_F^2 + \lambda\|\mathbf{X}\|_{2,1} \tag{3}$$

Then we can get the closed form solution of Eq.3 according to the following theorem:
Theorem 1:

$$\arg\min_{\mathbf{X}} p_{\mathbf{Z},L}(\mathbf{X}) = D_{\frac{\lambda}{L}}(\mathbf{Z} - \frac{1}{L}\nabla f_0(\mathbf{Z})), \tag{4}$$

where $D_\tau(.) :\ \mathbf{M} \in \mathbb{R}^{k\times k} \mapsto \mathbf{N} \in \mathbb{R}^{k\times k}$

$$\mathbf{N}_{i.} = \begin{cases} 0, & \|\mathbf{M}_{i.}\| \leq \tau; \\ (1-\tau/\|\mathbf{M}_{i.}\|)\mathbf{M}_{i.}, & \text{otherwise.} \end{cases} \tag{5}$$

We will derive it in the Appendix, and the whole algorithm is presented in Alg. 1.

Algorithm 1 Dictionary Selection

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

3.4. Sparse Reconstruction Cost using Weighted l_1 Minimization

This section details how to determine a testing sample $\mathbf{y}$ to be normal or not. As we mentioned in the previous subsection, the features of a normal frame can be linearly constructed by only a few bases in the dictionary $\mathbf{B}'$ while an abnormal frame cannot. A natural idea is to pursue a sparse representation and then use the reconstruction cost to judge the testing sample. In order to advance the accuracy of prediction, two more factors are considered here:

- In practice, the deformation or any un-predicated situation may happen to the video. Motivated by [25], we extend the dictionary from $\mathbf{B}'$ to $\Phi = [\mathbf{B}', \mathbf{I}_{m\times m}] \in \mathbb{R}^{m\times D}$, and $D = n+m$.
- If a basis in the dictionary appears frequently in the training dataset, then the cost to use it in the reconstruction should be lower, since it is a normal basis with high probability. Therefore, we design a weight matrix $\mathbf{W} = diag[w_1, w_2, ..., w_n, 1, ..., 1] \in \mathbb{R}^{D\times D}$ to capture this prior information. Each $w_i \in [0,1]$ corresponds to the cost of the i^{th} feature. For the artificial feature set $\mathbf{I}_{m\times m}$ in our new dictionary Φ, the cost for each feature is set to 1. The way to dynamically update $\mathbf{W}$ will be introduced in the following section.

Now, we are ready to formulate this sparse reforestation problem:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \Phi\mathbf{x}\|_2^2 \ +\ \lambda_1\|\mathbf{W}\mathbf{x}\|_1, \tag{6}$$

where $\mathbf{x} = [\mathbf{x}_0, \mathbf{e}_0]^T$, $\mathbf{x}_0 \in \mathbb{R}^n$, and $\mathbf{e}_0 \in \mathbb{R}^m$. This can be solved by linear programming using the interior-point method, which uses conjugate gradients algorithm to compute the optimized direction. Given a testing sample $\mathbf{y}$, we design a Sparsity Reconstruction Cost (SRC) using the minimal objective function value of Eq.6 to detect its abnormality:

$$S_w = \frac{1}{2}\|\mathbf{y} - \Phi\mathbf{x}^*\|_2^2 \ +\ \lambda_1\|\mathbf{W}\mathbf{x}^*\|_1. \tag{7}$$

A high SRC value implies a high reconstruction cost and a high probability of being an abnormal sample. In fact, the SRC function also can be equivalently mapped to the framework of Bayesian decision like in [11]. From a Bayesian view, the normal sample is the point with a higher probability, on the contrary the abnormal (outlier) sample is the point with a lower probability. We can estimate the normal

Algorithm 2 Abnormal Event Detection Framework

```
Input: Training dictionary $\Phi$, basis weight matrix $\mathbf{W}^0$, sequential input testing sample $\mathbf{Y} \in [\mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^T]$
Output: $\mathbf{W}$
for $t = 1, \cdots, \mathrm{T}$ do
    Pursuit the coefficient $\mathbf{x}^*$ by $l_1$ minimization:
        $\mathbf{x}^* = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}^t - \Phi\mathbf{x}\|_2^2 + \|\mathbf{W}^{t-1}\mathbf{x}\|_1$
    Calculate SRC function $S_w^t$ by Eq.7
    if $\mathbf{y}$ is normal then
        Select top K basis coefficients of $\mathbf{x}^*$
        Update $\mathbf{W}^t \longleftarrow \mathbf{W}^{t-1}$
    end if
end for
```

sample by maximizing the posteriori as follows:

$$\begin{aligned} \mathbf{x}^\star &= \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}, \Phi, \mathbf{W}) = \arg\max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}, \Phi, \mathbf{W}) p(\mathbf{x}|\Phi, \mathbf{W}) \\ &= \arg\max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}, \Phi) p(\mathbf{x}|\mathbf{W}) \\ &= \arg\min_{\mathbf{x}} - [\log p(\mathbf{y}|\mathbf{x}, \Phi) + \log p(\mathbf{x}|\mathbf{W})] \\ &= \arg\min_{\mathbf{x}} (\frac{1}{2}\|\mathbf{y} - \Phi\mathbf{x}\|_2^2 + \lambda_1 \|\mathbf{W}\mathbf{x}\|_1), \end{aligned} \tag{8}$$

where the first term is the likelihood $p(\mathbf{y}|\mathbf{x}, \Phi) \propto \exp(-\frac{1}{2}\|\mathbf{y} - \Phi\mathbf{x}\|_2^2)$, and the second term $p(\mathbf{x}; \mathbf{W}) \propto \exp(-\lambda_1 \|\mathbf{W}\mathbf{x}\|_1)$ is the prior distribution. This is consistent with our SRC function, as the abnormal samples correspond to smaller $p(\mathbf{y}|\mathbf{x}, \Phi)$, which results in greater SRC values.

3.5. Self-Updating

For a normal sample $\mathbf{y}$, we selectively update weight matrix $\mathbf{W}$ and dictionary Φ by choosing the top K bases with the highest positive coefficients of $\mathbf{x}_0^* \in \mathbb{R}^n$, and we denote the top K set as $\mathscr{S}_k = [s_1, \cdots, s_k]$.

As we have mentioned above, the contribution of each basis to the l_1 minimization reconstruction is not equal. In order to measure such a contribution, we use $\mathbf{W}$ to assign each basis a weight. The bases with higher weight, should be used more frequently and are more similarity to normal event and vice verse. We initialize $\mathbf{W}$ from matrix $\mathbf{X}$ of dictionary selection in Alg.1, i.e.,

$$\beta_i^0 = \|\mathbf{X}_{i.}\|_2, \quad w_i^0 = 1 - \frac{\beta_i^0}{\|\beta^0\|_1}, \tag{9}$$

where $\beta = [\beta_1, \ldots, \beta_n] \in \mathbb{R}^n$ denotes the accumulate coefficients of each basis, and $w_i \in [0, 1]$ (the smaller the value of w_i, the more likely a normal sample it is). The top K bases in $\mathbf{W}$ can be updated as follows:

$$\beta_i^{t+1} = \beta_i^t + \mathbf{x}_i^*, \{i \in \mathscr{S}_k\}, \quad w_i^{t+1} = 1 - \frac{\beta_i^{t+1}}{\|\beta^{t+1}\|_1}, \tag{10}$$

where $\mathscr{S}_k$ is the index set of the top K features in $\mathbf{W}$.

4. Experiments and Comparisons

To test the effectiveness of our proposed algorithm, we systematically apply it to several published datasets. The UMN dataset [1] is used to test the GAE; and the UCSD dataset [18] and the Subway dataset [2] are used to detect LAE. Moreover, we re-annotate the groundtruth of the Subway dataset using bounding boxes, where each box contains one abnormal event. Three different levels of measurements are applied for evaluation, which are Pixel-level, Frame-level and Event-level measurements.

4.1. Global Abnormal Event Detection

The UMN dataset consists of 3 different scenes of crowded escape events, and the total frame number is 7740 (1450, 4415 and 2145 for scenes $1-3$, respectively) with a 320×240 resolution. We initialize the training dictionary from the first 400 frames of each scene, and leave the others for testing. The type A basis in Fig.2(B), i.e., spatial basis, is used here. We split each image into 4×5 sub-regions, and extract the MHOF from each sub-region. We then concatenate them to build a basis with a dimension $m = 320$. Because the abnormal events cannot occur only in one frame, a temporal smooth is applied.

The results are shown in Fig.3, the normal/abnormal results are annotated as red/green color in the indicated bars respectively. In Fig.4, the ROC curves by frame-level measurement are shown to compare our SRC to three other measurements, which are

i. SRC with $\mathbf{W}$ as an identity matrix in Eq.7, $S = \frac{1}{2}\|\mathbf{y} - \Phi\mathbf{x}^*\|_2^2 + \lambda_1\|\mathbf{x}^*\|_1$.

ii. by formulating the sparse coefficient as a probability distribution, the entropy is used as a metric: $S_E = -\sum_i p_i \log p_i$, where $p(i) = |x(i)|/\|\mathrm{x}\|_1$, thus sparse coefficients will lead to a small entropy value.

iii. concentration function similar to [25], $S_S = T_k(\mathrm{x})/\|\mathrm{x}\|_1$, where $T_k(\mathrm{x})$ is the sum of the k largest positive coefficients of x (the greater the $\mathrm{S_s}$ the more likely a normal testing sample).

Moreover, Table 1 provides the quantitative comparisons to the state-of-the-art methods. The AUC of our method is from 0.964 to 0.995, which outperforms [20] and is comparable to [26]. However, our method is a more general solution, because it covers both LAE and GAE. Moreover, Nearest Neighbor (NN) method can also be used in high dimensional space by comparing the distances between the testing sample and each training samples. The AUC of NN is 0.93, which is lower than ours. This demonstrates the robustness of our sparse representation method over NN method.

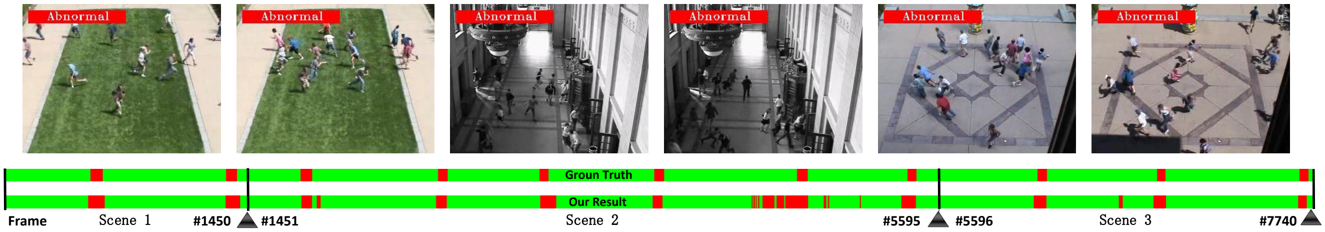


Figure 3. The qualitative results of the global abnormal event detection for three sample videos from UMN dataset. The top row represents snapshots of the result for a video in the dataset. At the bottom, the ground truth bar and the detection result bar show the labels of each frame for that video, where green color denotes the normal frames and red corresponds to abnormal frames.

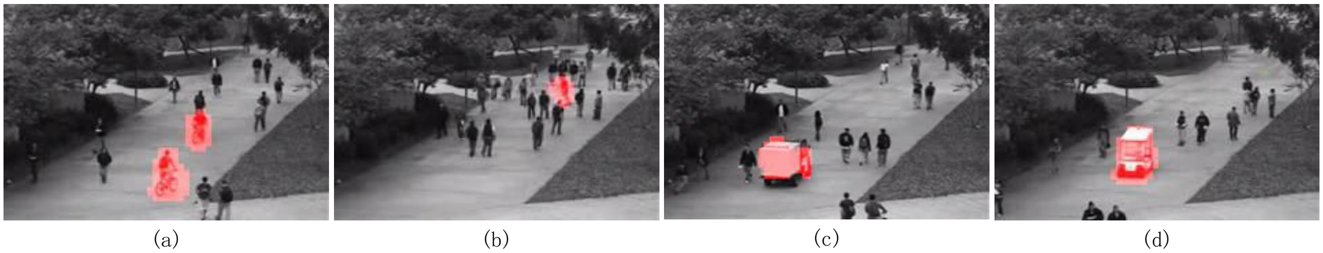


Figure 5. Examples of local abnormal event detections for UCSD Ped1 datasets. The objects, such as biker, skater and vehicle are all well detected.

Method	Area under ROC
Chaotic Invariants [26]	0.99
Social Force[20]	0.96
Optical flow [20]	0.84
NN	0.93
Ours Scene1	**0.995**
Ours Scene2	**0.975**
Ours Scene3	**0.964**

Table 1. The comparison of our proposed method with the state-of-the-art methods for GAE detection in the UMN dataset.

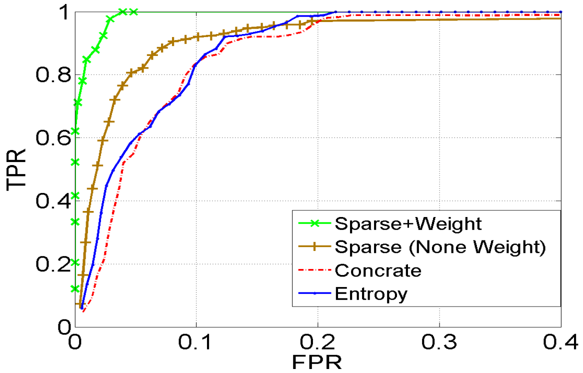


Figure 4. The ROCs for frame-level GAE detection in the UMN dataset. We compare different evaluation measurements, including SRC, SRC with $\mathbf{W} = \mathbf{I}$, concentration function S_S and entropy S_E. Our proposed SRC outperforms other measurements.

4.2. Local Abnormal Event Detection

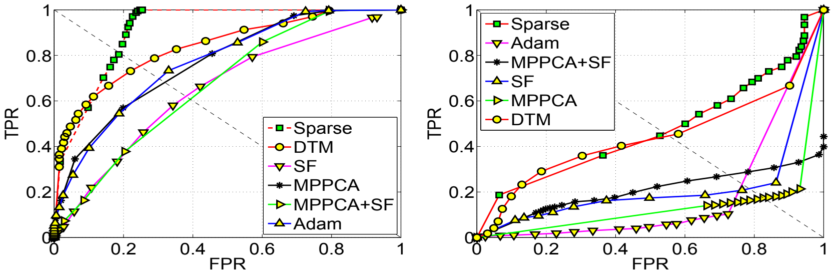
4.2.1 UCSD Ped1 Dataset

The UCSD Ped1 dataset contains pixel-level groundtruth. The training set contains 34 short clips for learning of normal patterns, and there is a subset of 10 clips in testing set provided with pixel-level binary masks, which identify the regions containing abnormal events. Each clip has 200 frames, with a 158×238 resolution. We split each frame into 7×7 local patches with 4-pixel overlapping. Type C basis in Fig.2(B), spatio-temporal basis, is selected to incorporate both local spatial and temporal information, with a dimension $m = 7 \times 16 = 102$. From each spatial location, we estimate a dictionary and use it to determine whether a testing sample is normal or not. A spatio-temporal smooth is adopted here to eliminate noise, which can be seen as a simplified version of spatio-temporal Markov Random Field [14].

Some image results are shown in Fig.5. Our algorithm can detect bikers, skaters, small cars, etc. In Fig.6, we compare our method with MDT, Social force and MPPCA, etc. Both pixel-level and frame-level measurements are defined in [18]. It is easy to find that our ROC curve outperforms others. In Fig.6(c), some evaluation results are presented: the Equal Error Rate (EER) (ours 19% $<$ 25%[18]), Rate of Detection (RD) (ours 46% $>$ 45%[18]) and Area Under Curve (AUC) (ours 46.1% $>$ 44.1%[18]), we can conclude that the performance of our algorithm outperforms the state-of-the-art methods.

4.2.2 Subway Dataset

The subway dataset is provided by Adam et al.[2], including two videos: "entrance gate" (1 hour 36 minutes long with 144249 frames) and "exit gate" (43 minutes long with 64900 frames). In our experiments, we resized the frames



(a) (b)

	EER	RD	AUC
SF [18]	31%	21%	17,9%
MPPCA [18]	40%	18%	20.5%
SF-MPPCA [18]	32%	18%	21.3%
MDT [18]	25%	45%	44.1%
Adam[2]	38%	24%	13.3%
Sparse	**19%**	**46%**	**46.1%**

(c)

Figure 6. The detection results of UCSD Ped1 dataset. (a) Frame-level ROCs for Ped1 Dataset, (b) Pixel-level ROCs for Ped1 Dataset, (c) Quantitative comparison of our method with [18][2]: EER is equal error rate; RD is rate of detection; and AUC is the area under ROC.

	Wrong Direction	No-Pay	Total	False Alarm
Ground truth	21/9	10/-	31/9	-/-
Adam[2]	17/9	-/-	17/9	4/2
Ours	**21/9**	**6/-**	**27/9**	**4/0**

Table 2. Comparisons of accuracy for subway videos. The first number in the slash (/) denotes the entrance gate result; the second is for the exit gate result.

from 512 × 384 to 320 × 240 and divided the new frames into 15 × 15 local patches with 6-pixel overlapping. The type B basis in Fig.2(B), temporal basis, is used with a dimension of $m = 16 \times 5 = 80$. The first 10 minutes are collected to estimate an optimal dictionary. The patch-level ROC curves for both data sets are presented in Fig. 8, where the positive detection and false positive correspond to each individual patch, and the AUCs are about 80% and 83%, respectively.

The examples of detection results are shown in Fig.7. In additional to wrong direction events, the no-payment events are also detected, which are very similar to normal “checking in” action. The event-level evaluation is shown in Table 2, our method detects all the wrong direction events, and also has a higher accuracy for no-payment events, comparing to others. This is because we use temporal basis which contains temporal causality context.

All experiments are run on a computer with 2GB RAM and a 2.6GHz CPU. The average computation time is 0.8s/frame for GAE, 3.8s/frame for UCSD dataset, and 4.6s/frame for the Subway dataset.

5. Conclusion

We propose a new criterion for abnormal event detection, namely the sparse reconstruction cost (SRC). Whether a testing sample is abnormal or not is determined by its sparse reconstruction cost, through a weighted linear reconstruction of the over-complete normal basis set. Thanks to the flexibility of our proposed dictionary selection model, our method cannot only support an efficient and robust estimation of SRC, but also easily handle both local abnormal events (LAE) and global abnormal events (GAE). By incrementally updating the dictionary, our method also supports online event detection. The experiments on three benchmark datasets show favorable results when compared with the state-of-the-art methods. Our method can also apply to other applications, such as event or action recognition.

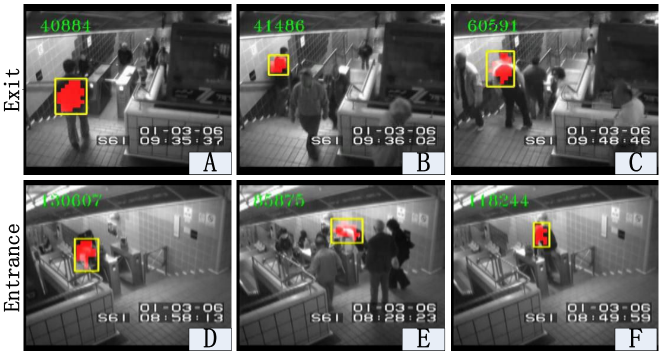


Figure 7. Examples of local abnormal events detection for Subway dataset. The top row and bottom row are from exit and entrance video sets, respectively, and red masks in the yellow rectangle indicate where the abnormality is, including wrong directions (A-D) and no-payments (E-F).

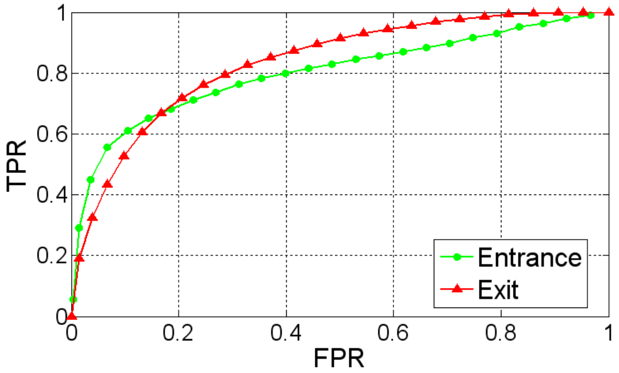


Figure 8. The frame-level ROC curves for both subway entrance and exit datasets

Acknowledgement

This work is supported in part by the Nanyang Assistant Professorship (SUG M58040015) to Dr. Junsong Yuan.

Appendix

We prove Theorem 1 here, where the optimization problem $\min_{\mathbf{X}} : \ p_{\mathbf{Z},L}(\mathbf{X})$ can be equivalently written as:

$$\begin{aligned} &\min_{\mathbf{X}} : \ f_0(\mathbf{Z}) + \langle \nabla f_0(\mathbf{Z}), \mathbf{X} - \mathbf{Z} \rangle + \frac{L}{2}\|\mathbf{X} - \mathbf{Z}\|_F^2 + \lambda \|\mathbf{X}\|_{2,1} \\ \Leftrightarrow &\min_{\mathbf{X}} : \ \frac{L}{2}\|(\mathbf{X} - \mathbf{Z}) + \frac{1}{L}\nabla f_0(\mathbf{Z}))\|_F^2 + \lambda \|\mathbf{X}\|_{2,1} \\ \Leftrightarrow &\min_{\mathbf{X}} : \ \frac{L}{2}\|\mathbf{X} - (\mathbf{Z} - \frac{1}{L}\nabla f_0(\mathbf{Z}))\|_F^2 + \lambda \|\mathbf{X}\|_{2,1} \\ \Leftrightarrow &\min_{\mathbf{X}} : \ \frac{L}{2}\|\mathbf{X} - (\mathbf{Z} - \frac{1}{L}\nabla f_0(\mathbf{Z}))\|_F^2 + \lambda \sum_{i=1}^{k} \|\mathbf{X}_{i.}\|_2 \end{aligned} \tag{11}$$

Since the l_2 norm is self dual, the problem above can be rewritten by introducing a dual variable $\mathbf{Y} \in \mathbb{R}^{k \times k}$:

$$\begin{aligned} &\min_{\mathbf{X}} : \ \frac{L}{2}\|\mathbf{X} - (\mathbf{Z} - \frac{1}{L}\nabla f_0(Z))\|_F^2 + \lambda \sum_{i=1}^{k} \max_{\|\mathbf{Y}_{i.}\|_2 \le 1} \langle \mathbf{Y}_{i.}, \mathbf{X}_{i.} \rangle \\ \Leftrightarrow &\max_{\|\mathbf{Y}_{i.}\|_2 \le 1} \min_{\mathbf{X}} : \ \frac{L}{2}\|\mathbf{X} - (\mathbf{Z} - \frac{1}{L}\nabla f_0(\mathbf{Z}))\|_F^2 + \lambda \sum_{i=1}^{k} \langle \mathbf{Y}, \mathbf{X} \rangle \\ \Leftrightarrow &\max_{\|\mathbf{Y}_{i.}\|_2 \le 1} \min_{\mathbf{X}} : \ \frac{1}{2}\|\mathbf{X} - (\mathbf{Z} - \frac{1}{L}\nabla f_0(\mathbf{Z}) - \frac{\lambda}{L}\mathbf{Y})\|_F^2 \\ &- \frac{1}{2}\|\mathbf{Z} - \frac{1}{L}\nabla f_0(\mathbf{Z}) - \frac{\lambda}{L}\mathbf{Y}\|_F^2 \end{aligned} \tag{12}$$

The second equation is obtained by swapping “max” and “min”. Since the function is convex with respect to **X** and concave with respect to **Y**, this swapping does not change the problem by the Von Neumann minimax theorem. Letting $\mathbf{X} = \mathbf{Z} - \frac{1}{L}\nabla f_0(\mathbf{Z}) - \frac{\lambda}{L}\mathbf{Y}$, we obtain an equivalent problem from the last equation above

$$\max_{\|\mathbf{Y}_{i.}\|_2 \le 1} : \ -\frac{1}{2}\|\mathbf{Z} - \frac{1}{L}\nabla f_0(\mathbf{Z}) - \frac{\lambda}{L}\mathbf{Y}\|_F^2 \tag{13}$$

Using the same substitution as above, $\mathbf{Y} = -\frac{L}{\lambda}(\mathbf{X} - \mathbf{Z} + \frac{1}{L}\nabla f_0(\mathbf{Z}))$, we change it into a problem in terms of the original variable **X** as

$$\begin{aligned} &\min_{\|\frac{L}{\lambda}(\mathbf{X} - \mathbf{Z} + \frac{1}{L}\nabla f_0(\mathbf{Z}))_{i.}\|_2 \le 1} : \ \|\mathbf{X}\|_F^2 \\ \Leftrightarrow &\sum_{i=1}^{k} \min_{\|\mathbf{X}_{i.} - (\mathbf{Z} - \frac{1}{L}\nabla f_0(\mathbf{Z}))_{i.}\|_2 \le \frac{\lambda}{L}} : \ \|\mathbf{X}_{i.}\|_2^2. \end{aligned} \tag{14}$$

Therefore, the optimal solution of the first problem in Eq. 14 is equivalent to the last problem in Eq. 14. Actually, each row of **X** can be optimized independently in the last problem. Considering each row of **X** respectively, we can get the closed form as $\arg\min_{\mathbf{X}} p_{\mathbf{Z},L}(\mathbf{X}) = D_{\frac{\lambda}{L}}(\mathbf{Z} - \frac{1}{L}\nabla f_0(\mathbf{Z}))$.

References

[1] Unusual crowd activity dataset of University of Minnesota,from http://mha.cs.umn.edu/movies/crowdactivity-all.avi.

[2] E. S. I. R. D. Adam, A.; Rivlin. Robust real-time unusual event detection using multiple fixed-location monitors. *TPAMI*, 30(3)Volume 30:555 – 560, 2008.

[3] S. Avidan. Ensemble tracking. *IEEE transactions on pattern analysis and machine intelligence*, pages 261–271, 2007.

[4] Y. Benezeth, P. Jodoin, V. Saligrama, and C. Rosenberger. Abnormal events detection based on spatio-temporal co-occurences. In *CVPR*, 2009.

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

[6] O. Boiman and M. Irani. Detecting irregularities in images and in video. *IJCV*, 74(1):17–31, 2007.

[7] Y. Cong, H. Gong, S. Zhu, and Y. Tang. Flow mosaicking: Real-time pedestrian counting without scene-specific learning. In *CVPR*, pages 1093–1100, 2009.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[9] P. D.Helbing. Social force model for pedestrian dynamics. *Physical Review*, E, 51:4282, 1995.

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

[11] K. Huang and S. Aviyente. Sparse representation for signal classification. In *NIPS*, 2007.

[12] L. Itti and P. Baldi. A principled approach to detecting surprising events in video. In *CVPR*, 2005.

[13] 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.

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

[15] L. Kratz and K. Nishino. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 2009.

[16] C. Liu, W. Freeman, E. Adelson, and Y. Weiss. Human-assisted motion annotation. In *CVPR*, 2008.

[17] J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. In *ICCV*, 2009.

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

[19] Y. Nesterov. *Gradient methods for minimizing composite objective function*. CORE, 2007.

[20] M. S. Ramin Mehran, Alexis Oyama. Abnormal crowd behavior detection using social force model. In *CVPR*, 2009.

[21] M. S. Saad Ali. Floor fields for tracking in high density crowd scenes. In *ECCV*, 2008.

[22] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, 2002.

[23] I. Tziakos, A. Cavallaro, and L. Xu. Event monitoring via local motion abnormality detection in non-linear subspace. *Neurocomputing*, 2010.

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

[25] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *TPAMI*, 31(2):210–227, 2008.

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

[27] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *CVPR*, pages 2442–2449, 2009.

[28] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

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