Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Online Detection of Abnormal Events Using Incremental Coding Length

Jayanta K. Dutta and **Bonny Banerjee**
Institute for Intelligent Systems, and Department of Electrical & Computer Engineering
The University of Memphis
Memphis, TN 38152, USA
{jkdutta, bbnerjee}@memphis.edu

Abstract

We present an unsupervised approach for abnormal event detection in videos. We propose, given a dictionary of features learned from local spatiotemporal cuboids using the sparse coding objective, the abnormality of an event depends jointly on two factors: the frequency of each feature in reconstructing all events (or, rarity of a feature) and the strength by which it is used in reconstructing the current event (or, the absolute coefficient). The Incremental Coding Length (ICL) of a feature is a measure of its entropy gain. Given a dictionary, the ICL computation does not involve any parameter, is computationally efficient and has been used for saliency detection in images with impressive results. In this paper, the rarity of a dictionary feature is learned online as its average energy, a function of its ICL. The proposed approach is applicable to real world streaming videos. Experiments on three benchmark datasets and evaluations in comparison with a number of mainstream algorithms show that the approach is comparable to the state-of-the-art.

Introduction

Video cameras that monitor round-the-clock are ubiquitous and expanding their reach exponentially. Due to the sheer amount of data these sensors can generate, the resources required to store, protect personal information, and analyze them are enormous. Since anomalous events rarely occur, it is imperative for smart sensors to detect such events which may then be stored and decisions can be taken.

An anomalous event may be defined as one that stands out due to some property that occurs infrequently among the events in its neighboring spatiotemporal locations. Anomaly detection can be posed as an outlier detection problem where normal events are modeled and anomaly is detected as significant deviation from the norm (Hou and Zhang 2008; Cong, Yuan, and Liu 2013; Roshtkhari and Levine 2013; Zhao, Fei-Fei, and Xing 2011; Borji and Itti 2012). In case of spatiotemporal data, such as videos, it is necessary to detect the abnormal frames (i.e. temporal anomaly) and to localize the abnormal events within the detected frames (i.e. spatial anomaly). The problem of anomalous event detection in videos is difficult due to three primary reasons. First, the data is unlabeled, so supervised training for binary classification (anomalous vs. not) is not an option. Second, the data size is enormous, so storing all the data is not feasible. Finally, the underlying data distribution is non-stationary and involves concept drifts; consequently normality may vary over time.

Our goal is to build a fast and accurate abnormal event detection algorithm for practical real-world applications. It is desirable that a method quickly builds a model of normality and detects anomalies while incrementally updating itself in an unsupervised manner as new patterns are observed. The contributions of this paper are twofold.

A new definition of anomaly. Given a dictionary of features learned from local spatiotemporal cuboids using the sparse coding objective (Olshausen and Field 1996; Mairal et al. 2010), we define a data point as anomalous if it consists of one or more features with significant strength that rarely occur in the other observed data points. Formally, given a set of data points $\mathbf{X} = \{\vec{x}_i\}_{i=1}^{N}$, an anomalous data point $\vec{x}_c$ is defined as:

$$\vec{x}_c \in \mathbf{X}, \quad \zeta(\vec{x}_c|\mathbf{X}) \geq \omega \tag{1}$$

where $\vec{x}_i \in \mathbb{R}^m$, m is the dimension of each data point, $\zeta : \mathbf{X} \rightarrow \mathbb{R}^+$ is a scoring function, $\mathbb{R}^+$ is the set of non-negative real numbers and ω is a threshold. ζ assigns an *outlier score* to each data in $\mathbf{X}$ based on its frequency of occurrence in comparison to that of the other observed data. Rarer data points are assigned higher scores. A data could be a pixel, region, object or event, depending on the nature of the data and the goal. For example, if the data is a video sequence, $\mathbf{X}$ is a set of events $\vec{x}_i$ defined at each point (x_i, y_i, t_i) in the video where (x_i, y_i) refers to the spatial location in a frame and t_i is the index of the frame. The crux of the problem is to discover the function ζ such that unusual or rare data in a dataset can be detected.

Evaluating ICL for scoring spatiotemporal data points for anomaly detection. The Incremental Coding Length (ICL) of a feature is a measure of its entropy gain. Given a dictionary of features, the ICL computation does not involve any parameter, is computationally efficient and has been used for saliency detection in images with impressive results (Hou and Zhang 2008). In this paper, the rarity of a spatiotemporal feature in the dictionary is learned online as its average energy, a function of its ICL. No prior assumption is made regarding the data or nature of anomaly. Due to its unsupervised and online operation, the proposed approach is appli-

cable to real world streaming videos. Experiments on three benchmark datasets and evaluations in comparison with a number of mainstream algorithms show that the approach is comparable to the state-of-the-art.

Rest of this paper is organized as follows. In the next section, prior related work is reviewed. Then the proposed framework is described followed by experimental results on a number of benchmark datasets and comparison with a number of mainstream anomaly detection algorithms. Finally, the paper ends with conclusions.

Related Work

Typically, anomalous event detection involves three operations: preprocessing of input data to extract low-level representation, learning abstractions of this representation to extract mid-level features, and scoring the data points with reference to these features to detect anomalies. Execution order of these operations often depends on whether the approach is online or not. In case of videos, it is common to represent a spatiotemporal cuboid of fixed size as a data point. Examples of preprocessing methods include histogram of optical flow (Adam et al. 2008), spatiotemporal gradient (Kratz and Nishino 2009), social force model (Mehran, Oyama, and Shah 2009), chaotic invariant (Wu, Moore, and Shah 2010). Mid-level features have been learned using different methods, such as dimensionality reduction (e.g. PCA, ICA, clustering), sparse coding (Zhao, Fei-Fei, and Xing 2011; Cong, Yuan, and Liu 2013; Lu, Shi, and Jia 2013), Gaussian mixture model (Kim and Grauman 2009), mixtures of dynamic textures (Li, Mahadevan, and Vasconcelos 2014), hidden Markov model (Kratz and Nishino 2009), Markov random field (Benezeth et al. 2009), latent Dirichlet allocation (Wu, Moore, and Shah 2010). A hierarchy of mid-level features can also be learned using deep learning. A number of methods for scoring the data points have been explored, including reconstruction error (Zhao, Fei-Fei, and Xing 2011; Cong, Yuan, and Liu 2013; Lu, Shi, and Jia 2013), prediction error (Banerjee and Dutta 2013a; 2013b; 2014), rarity index (Hou and Zhang 2008; Borji and Itti 2012), information content (Li, Mahadevan, and Vasconcelos 2014), and density-based scoring (Wu, Moore, and Shah 2010; Kim and Grauman 2009).

In particular, three approaches reported in the literature bear resemblance to our proposed approach. Zhao, Fei-Fei, and Xing (2011) preprocessed cuboids at interest points to compute histograms of gradient and optical flow. A dictionary is learned by minimizing the reconstruction error regularized by sparsity and smoothness constraints. The saliency score of an event is the minimum value of this objective function. Cong, Yuan, and Liu (2013) preprocessed cuboids to extract multi-scale histogram of optical flow (MHOF) features. A subset of input cuboids are selected by minimizing the reconstruction error with sparsity constraint to form a dictionary. Each dictionary element is assigned a cost based on its frequency of occurrence. The saliency score of an event is its sparse reconstruction cost. Lu, Shi, and Jia (2013) used cuboids at multiple spatial scales and learned multiple dictionaries. The saliency score of an event is the minimum reconstruction cost over all the dictionaries.

Proposed Framework

The idea behind the proposed framework is as follows. First, a video is divided into an ensemble of maximally-overlapping clips of M frames and each clip is processed in an online manner. Then local spatiotemporal volumes are extracted and represented as a sparse linear combination of dictionary features. Abnormal events can be defined in terms of rarity of the features. Rarely occurring features are more abnormal than the frequently occurring ones. Each feature is assigned an ICL score, which is defined as the ensemble's entropy gain during the activity increment of corresponding feature (Hou and Zhang 2008). Finally, anomaly is obtained by the weighted summation of the absolute activity by the energy of all features.

Video Representation

The proposed framework uses local spatiotemporal volumes around detected interest points in each clip as an input representation. Here, we adopt a spatiotemporal interest point detector (Dollár et al. 2005) to extract cuboids which contain the spatiotemporally windowed pixels. Before learning the dictionary, each cuboid is converted to a vector and normalized to have a unit ℓ_2 norm. The proposed framework can also be applied over other video descriptors. Figure 1 shows some of the frames from different datasets with detected spatiotemporal interest points within each frame.

Online Sparse Dictionary Learning

Notation. In this paper, matrices are denoted by bold uppercase letters, while lowercase letters with vector sign denote column vectors. The columns of a matrix are represented by corresponding lowercase letters. The elements of a vector are denoted by letters without vector sign. $\vec{0}$ denotes a zero vector with size depending on the context. The elements of a matrix are denoted using subscript where row and column indices are separated by a comma. Time indices are denoted in a parenthesis after the variable.

Let $\vec{x} \in \mathbb{R}^m$ be a data point. It admits a sparse representation $\vec{\gamma} \in \mathbb{R}^k$ over a dictionary of k features, $\mathbf{D} \in \mathbb{R}^{m \times k}$, if $\vec{x}$ can be represented as a linear combination of κ features in $\mathbf{D}$ and $\kappa \ll k$.

The dictionary learning task is to train a dictionary such that it is well adapted for reconstructing a set of data points. Given a set of N data points $\mathbf{X} = [\vec{x}_1, ..., \vec{x}_N]$, a dictionary of k features with a sparsity constraint can be learned by solving the following optimization problem:

$$\min_{\mathbf{\Gamma}, \mathbf{D}} \frac{1}{2} \sum_{i=1}^{N} \|\vec{x}_i - \mathbf{D}\vec{\gamma}_i\|_2^2 \text{ subject to } \|\vec{\gamma}_i\|_0 \leq \kappa \, \forall i \quad (2)$$

where $\mathbf{\Gamma} = [\vec{\gamma}_1, ..., \vec{\gamma}_N] \in \mathbb{R}^{k \times N}$ is a sparse representation matrix, $\|.\|_0$ denotes the ℓ_0 pseudo-norm, the number of non-zero elements. κ is the maximum number of non-zero elements allowed in each $\vec{\gamma}_i$ and $\kappa \ll k$. Each element $\vec{d}_j \in \mathbf{D}$ $(j = 1, ..., k)$ is constrained to have a unit ℓ_2 norm.

An online dictionary learning algorithm draws one input, $\vec{x}(t)$, or a small batch of inputs, $\mathbf{X}(t) = [\vec{x}_1(t), ..., \vec{x}_n(t)] \in$

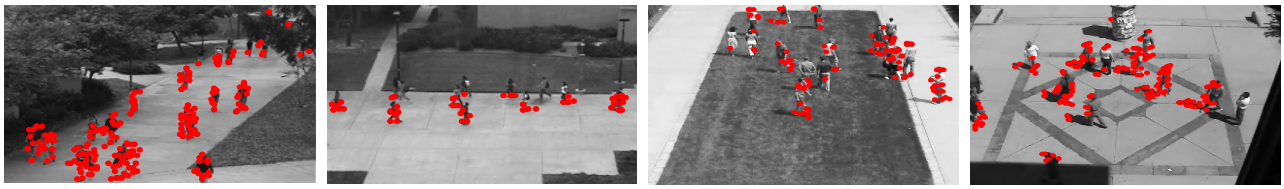

Figure 1: Examples of detected spatiotemporal interest points (best viewed in color) using the method in (Dollár et al. 2005).

$\mathbb{R}^{m\times n}$, at any time t, followed by two steps: sparse coding and dictionary update.

Sparse coding: Given a fixed dictionary $\mathbf{D} \in \mathbb{R}^{m\times k}$ and a data point $\vec{x} \in \mathbb{R}^m$, the sparse linear representation $\vec{\gamma} \in \mathbb{R}^k$ is obtained by solving the following sparse approximation problem:

$$\min_{\vec{\gamma}} \frac{1}{2}\|\vec{x} - \mathbf{D}\vec{\gamma}\|_2^2 \quad \text{subject to} \quad \|\vec{\gamma}\|_0 \leq \kappa \tag{3}$$

This sparse approximation problem can be efficiently solved using Orthogonal Matching Pursuit (OMP) (Pati, Rezaiifar, and Krishnaprasad 1993) which is a greedy forward selection algorithm. Since at any instant t, multiple cuboids can be extracted from the detected interest points within the clip, we use the Batch-OMP (Rubinstein, Zibulevsky, and Elad 2008), which speeds up the process considerably.

Dictionary update: At any time t, the optimal dictionary is the solution to the following optimization problem:

$$\begin{aligned}\mathbf{D}(t) &= \underset{\mathbf{D}}{\operatorname{argmin}} \frac{1}{t}\sum_{i=1}^{t}\left(\frac{1}{2}\|\vec{x}(t) - \mathbf{D}\vec{\gamma}(t)\|_2^2\right) \\ &= \underset{\mathbf{D}}{\operatorname{argmin}} \frac{1}{t}\left(\frac{1}{2}\mathrm{Tr}(\mathbf{D}^T\mathbf{D}\mathbf{A}(t)) - \mathrm{Tr}(\mathbf{D}^T\mathbf{B}(t))\right)\end{aligned}$$

This paper follows the online dictionary update algorithm reported in (Mairal et al. 2010) which uses block-coordinate descent with warm restarts (Bertsekas 1999) for updating the dictionary. This procedure reduces the memory requirements by storing only the matrices $\mathbf{A}(t) = \sum_{h=1}^{t}\vec{\gamma}(h)(\vec{\gamma}(h))^T \in \mathbb{R}^{k\times k}$ and $\mathbf{B}(t) = \sum_{h=1}^{t}\vec{x}(h)(\vec{\gamma}(h))^T \in \mathbb{R}^{m\times k}$ rather than storing all the input signals and their sparse representations. Mini-batch extension of this algorithm is used here, as we have multiple cuboids at any time. It has been shown that the mini-batch extension improves the convergence speed of this online dictionary update algorithm (Mairal et al. 2010). The matrices are initialized to zeros and are updated as follows:

$$\mathbf{A}(t) = \mathbf{A}(t-1) + \frac{1}{n}\sum_{i=1}^{n}\vec{\gamma}_i(t)(\vec{\gamma}_i(t))^T \tag{4}$$

$$\mathbf{B}(t) = \mathbf{B}(t-1) + \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i(t)(\vec{\gamma}_i(t))^T \tag{5}$$

The pseudo-code for dictionary update is presented in Algorithm 1.

Algorithm 1 Online Dictionary Update

1: Input: Dictionary $\mathbf{D} = [\vec{d}_1, ..., \vec{d}_k] \in \mathbb{R}^{m\times k}$, $\mathbf{A} = [\vec{a}_1, ..., \vec{a}_k] \in \mathbb{R}^{k\times k}$, $\mathbf{B} = [\vec{b}_1, ..., \vec{b}_k] \in \mathbb{R}^{m\times k}$
2: Output: Updated dictionary $\mathbf{D} \in \mathbb{R}^{m\times k}$
3: **while** convergence criterion not met **do**
4: **for** $j = 1, 2, ..., k$ **do**
5: $\vec{u}_j \leftarrow \frac{1}{\mathbf{A}[j,j]}(\vec{b}_j - \mathbf{D}\vec{a}_j) + \vec{d}_j$
6: $\vec{d}_j \leftarrow \frac{1}{\|\vec{u}_j\|_2}\vec{u}_j$
7: **end for**
8: **end while**

Abnormal Event Detection

The rarity of dictionary features is computed using ICL (Hou and Zhang 2008). At any time t, the sparse linear coefficient vectors $\mathbf{\Gamma} = [\vec{\gamma}_1, ..., \vec{\gamma}_n] \in \mathbb{R}^{k\times n}$ for input cuboids, $\mathbf{X}(t) = [\vec{x}_1(t), ..., \vec{x}_n(t)] \in \mathbb{R}^{m\times n}$ with respect to a learned dictionary $\mathbf{D}$ are computed using Batch-OMP. The activity ratio, $p_j(t)$ for j^{th} feature over the inputs at time t is computed as:

$$p_j(t) = \frac{\sum_{h=1}^{n}|\mathbf{\Gamma}_{j,h}(t)|}{\sum_{i=1}^{k}\sum_{h=1}^{n}|\mathbf{\Gamma}_{i,h}(t)|} \tag{6}$$

The summary activity ratio $\vec{q}(t)$ at any time t be incrementally updated as follows:

$$\vec{q}(t) = (1-\alpha(t))\vec{q}(t-1) + \alpha(t)\vec{p}(t) \tag{7}$$

where $\alpha(t)$ is a parameter, a function of time, and $0 < \alpha(t) \leq 1\ \forall t$. Thus, in order to determine the new estimate $\vec{q}(t)$, the prior estimate $\vec{q}(t-1)$ is weighted by $1-\alpha(t)$ while the new outcome $\vec{p}(t)$ is weighted by $\alpha(t)$. If $\alpha(t) = 1/t$, $\vec{q}(t)$ is the mean of the activity ratio since the beginning of time. If $\alpha(t) = 1/t_1$ where t_1 is a constant, a positive integer, $\vec{q}$ is a soft moving average of the activity ratio for the last t_1 time instants. It does not discard everything before the last t_1 instants but assigns them much less weight in the estimation process. The latter case is particularly useful if the data distribution changes over time. The summary activity ratio of each dictionary atom is initialized to $1/k$. It is obvious because at the beginning, all the dictionary atoms are equally likely to be used for reconstruction.

Given the summary activity ratio at any time t, ICL can be defined as (Hou and Zhang 2008):

$$\begin{aligned}\text{ICL}(q_j) &= \frac{\partial H(\vec{q})}{\partial q_j} \\ &= -H(\vec{q}) - q_j - \log q_j - q_j \log q_j\end{aligned}$$

After calculating ICL score, salient feature set at any time t is defined as: $S(t) = \{j \mid \text{ICL}(q_j(t)) > 0\}$. The amount of energy received by each salient feature at any time t, $\theta_j(t)$ where $j \in S(t)$ is calculated as:

$$\theta_j(t) = \begin{cases} \dfrac{\text{ICL}(q_j(t))}{\sum_{i \in S(t)} \text{ICL}(q_i(t))}, & \text{if } j \in S(t) \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

The value of $\theta_j(t)$ indicates how rarely the j^{th} feature is used to represent an input signal in some temporal neighborhood, the size of the neighborhood is determined by $\alpha(t)$. For a suitable $\alpha(t)$, this value will be high for rarely used features and low for the features used very often. Finally, given a data point $\vec{x}$ and its coefficient vector $\vec{\gamma}$ with respect to a learned dictionary $\mathbf{D}$, the anomaly score g is defined as:

$$\zeta(\vec{x}) = g = |\vec{\gamma}|^T \vec{\theta} \quad (9)$$

Here, sampling is sparse in space but dense in time. An anomaly map is generated for the current frame (except at the temporal boundaries) taking into account all frames in the temporal window of which the current frame is the last. Each cuboid, extracted by a spatiotemporal interest point detector (Dollár et al. 2005), receives an anomaly score which is shared by all pixels in the cuboid in the current frame. This image of scores is then blurred with an appropriate Gaussian filter to generate the final anomaly map for the current frame.

Experimental Results

The proposed approach is evaluated on a number of benchmark datasets using multiple evaluation criteria. No prior assumption is made regarding the data or nature of anomaly. Results obtained from this extensive experimentation show that the proposed rarity-based approach is comparable to the state-of-the-art.

Datasets

The UCSD, UMN and Subway are among the most commonly used benchmark datasets for abnormal event detection in videos.

UCSD dataset: The UCSD dataset (Mahadevan et al. 2010) is organized into two sub-datasets, Ped1 and Ped2. Ped1 contains 34 training and 36 testing video clips of 158×238 pixel resolution. Ped2 contains 16 training and 12 testing video clips of 240×360 pixel resolution. The training sets have all normal events and contain only pedestrians. Each testing video clip contains at least one abnormal event with the presence of non-pedestrians on the walk-way, such as bicyclists, skaters, small cars, and people in wheelchairs.

UMN dataset: The UMN dataset (Mehran, Oyama, and Shah 2009) contains three different crowded scenes. Normal events consists of individuals wandering around or organized in groups. Abnormal events consist of crowd escape scenes. The total number of frames is 7740 (1450, 4415 and 2145 for scenes 1, 2 and 3 respectively) with a 320×240 pixel resolution.

Subway dataset: The Subway dataset (Adam et al. 2008) includes two videos: entrance gate (1 hour 36 minutes in duration consisting of 144,249 frames) and exit gate (43 minutes in duration consisting of 64,900 frames) with a 512×384 pixel resolution. Normal behaviors include people entering and exiting the station while abnormal events include people moving in the wrong direction, such as exiting the entrance and entering the exit, or avoiding payment.

Evaluation Criteria

Following (Cong, Yuan, and Liu 2013; Li, Mahadevan, and Vasconcelos 2014), two criteria were used for evaluation of abnormal event detection accuracy: frame level criterion and pixel level criterion.

- **Frame-level criterion:** An algorithm determines the frames that contain abnormal events. The result is compared to the frame-level ground truth annotation of each frame and the number of true and false positive frames are calculated.
- **Pixel-level criterion:** An algorithm determines the pixels that are related to abnormal events. If at least 40% of the truly anomalous pixels are detected for an abnormal frame, it is considered a correct detection.

For both the cases, true positive rate (TPR) is calculated as the ratio between number of true positive frames and number of positive frames, and false positive rate (FPR) is calculated as the ratio between number of false positive frames and number of negative frames. TPR and FPR is calculated for different threshold values in the range from minimum to maximum anomaly score. Then, ROC curve is drawn as the TPR vs. FPR. Finally, the performance is summarized using *equal error rate* (EER) for frame level criterion and *rate of detection* (RD) for pixel level criterion (Li, Mahadevan, and Vasconcelos 2014). A low EER value and high RD value indicates a better performance.

Metrics for comparison depend on the type of available ground truth. Only frame level ground truth is available for UMN dataset, hence we are only able to compute EER and AUC. Only event level ground truth is available for Subway dataset, hence we are only able to compute the event level detection accuracy. Ground truths for both frame and pixel level criteria are available for UCSD datasets, hence we are able to compute both, EER and RD.

Performance Evaluation

The proposed method was tested on the described datasets and compared with a number of state-of-the art methods, including H-MDT (Li, Mahadevan, and Vasconcelos 2014), Sparse (Cong, Yuan, and Liu 2013), STC (Roshtkhari and Levine 2013), MPPCA (Kim and Grauman 2009), Social Force (Mehran, Oyama, and Shah 2009), LMH (Adam et al.

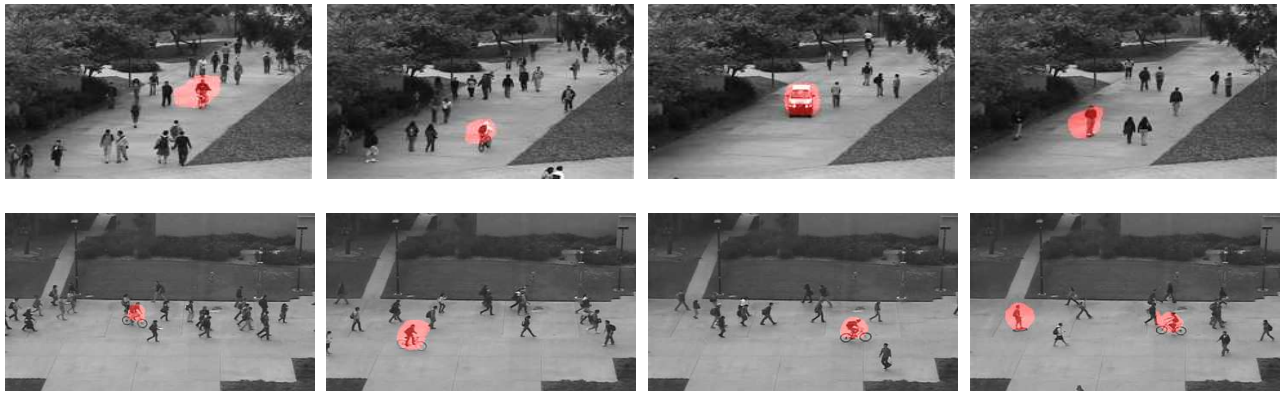

Figure 2: Abnormal frames and their detection result from UCSD Ped1 (top row) and UCSD Ped2 (bottom row) using our model. The bikers, skaters and cars were detected as anomalous patterns (highlighted in red, best viewed in color). The proposed method can detect multiple anomalous patterns within a single frame.

Figure 3: Abnormal frames from UMN dataset. Anomalous regions, as detected by our model, are highlighted in red (best viewed in color).

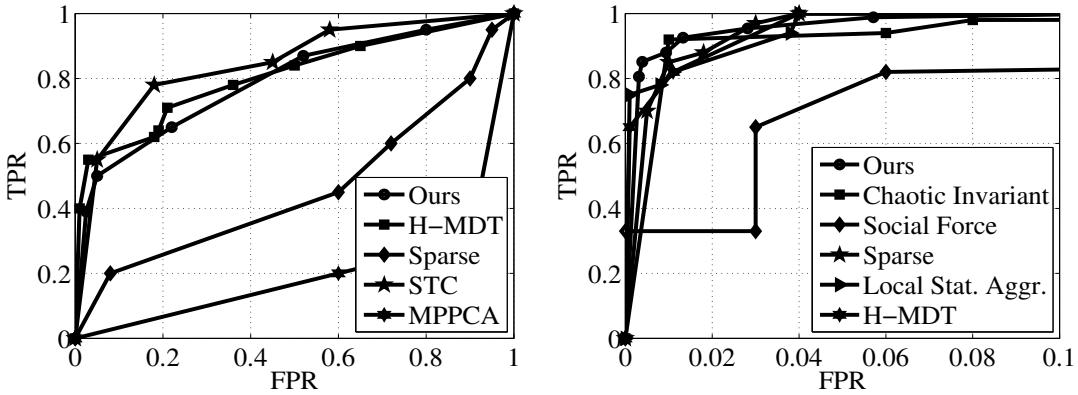

Figure 4: (a) ROC curves for pixel-level criterion on UCSD Ped1 dataset. (b) ROC curves for frame-level criterion on UMN dataset.

2008), Chaotic Invariant (Wu, Moore, and Shah 2010), Local Statistical Aggregates (Saligrama and Chen 2012), Dynamic SC (Zhao, Fei-Fei, and Xing 2011), and SCL (Lu, Shi, and Jia 2013). In our experiments, the videos were presented randomly and the frames in each video were presented sequentially in order.

UCSD dataset: The entire test set was used for the evaluation of our model on Ped1 and Ped2 datasets. The anomaly maps are generated based on the input cuboids of size $13 \times 13 \times 10$ pixels. The proposed model could detect bikers, skaters, small cars as abnormal events. It could also detect multiple anomalous patterns within a single frame. A few of the results are shown in Figure 2. The ROC curves in Figure 4a and Table 1 compare our model with others'. These comparisons reveal that the performance of our model is at par with the state-of-the-art.

UMN dataset: For the UMN dataset, the first 400 frames of each scene were used to learn the dictionary and energy for each dictionary feature. The other frames were used for testing. Figure 3 shows a few of the results. A comparison with

Figure 5: Anomaly detection in Subway dataset. Top row represents entrance gate and bottom row represents exit gate. Anomaly detection includes detection of wrong direction events and no payment events. Anomalies detected by our model are highlighted in red (best viewed in color).

Table 1: Anomaly detection performance on USCD Ped1 and Ped2 datasets.

Method	EER (Ped1)	RD (Ped1)	EER (Ped2)	RD (Ped2)
Ours	19.8	69.5	22.3	67.5
H-MDT	17.8	75	18.5	70
Sparse	19	46	X	X
STC	15	73	13	74
MPPCA	35.6	23.2	35.8	22.4
Social Force	36.5	40.9	35	27.6
LMH	38.9	32.6	45.8	22.4

prior results reported in the literature, under the frame-level criterion, is presented in Figure 4b and Table 2. Performance of our model is comparable to the state-of-the-art in the literature.

Table 2: Quantitative comparison between different methods on UMN dataset.

Method	AUC	EER
Ours	99.5	3.65
Chaotic Invariant	99.4	5.3
Sparse	99.6	2.8
Social Force	94.9	12.6
Local Statistical Aggregates	99.5	3.4
H-MDT	99.5	3.7

Subway dataset: For the Subway dataset, each frame was resized to 320×240 pixel resolution. The dictionary was learned using cuboids of size $13 \times 13 \times 10$ pixels. The first ten minutes of each video were used to learn the dictionary and energy for each dictionary features. Figure 5 shows a few detection results which include detection of wrong direction events as well as no payment events. Table 3 compares the performance of our model with other existing models. It shows that our model is comparable to the state-of-the-art models reported in the literature.

Table 3: Performance analysis on the Subway dataset.

Method	Dataset	Abnormal events	False alarm
Ours	Entrance	60/66	5
	Exit	19/19	2
STC	Entrance	61/66	4
	Exit	19/19	2
MPPCA	Entrance	57/66	6
	Exit	19/19	3
Dynamic SC	Entrance	60/66	5
	Exit	19/19	2
Sparse	Entrance	27/31	4
	Exit	9/9	0
SCL	Entrance	57/66	4
	Exit	19/19	2

Conclusions

A rarity-based approach for anomaly detection in streaming videos was presented. Given a dictionary of features learned from local spatiotemporal cuboids using the sparse coding objective, we define a data point as anomalous if it consists of one or more features with significant strength that rarely occur in the other observed data points. The rarity of each feature was approximated online using ICL. The anomaly score for an input data was computed as the sum, over all features, of the average energy multiplied by absolute coefficients. Finally, the anomaly map was generated from this image of anomaly scores.

The proposed approach was extensively experimented with a number of benchmark datasets and evaluated in comparison to a number of mainstream algorithms. No prior assumption was made regarding the data or nature of anomaly. The unsupervised and online operation of the proposed method allows it to deal with space- and time-varying data and is useful to real-time applications. Experimental results, reported in this paper, showed that the proposed rarity-based approach is comparable to the state-of-the-art.

Acknowledgements

This work was partially supported by the U.S. National Science Foundation under CISE Grant no. 1231620.

References

Adam, A.; Rivlin, E.; Shimshoni, I.; and Reinitz, D. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *IEEE TPAMI* 30(3):555–560.

Banerjee, B., and Dutta, J. K. 2013a. Efficient learning from explanation of prediction errors in streaming data. In *IEEE International Conference on Big Data*, 14–20. IEEE.

Banerjee, B., and Dutta, J. K. 2013b. A predictive coding framework for learning to predict changes in streaming data. In *IEEE 13th International Conference on Data Mining Workshops (ICDMW)*, 497–504.

Banerjee, B., and Dutta, J. K. 2014. SELP: A general-purpose framework for learning the norms from saliencies in spatiotemporal data. *Neurocomputing: Special Issue on Brain Inspired Models of Cognitive Memory* 138:41–60.

Benezeth, Y.; Jodoin, P.; Saligrama, V.; and Rosenberger, C. 2009. Abnormal events detection based on spatio-temporal co-occurences. In *CVPR*, 2458–2465. IEEE.

Bertsekas, D. P. 1999. *Nonlinear programming*. Athena Scientific Belmont.

Borji, A., and Itti, L. 2012. Exploiting local and global patch rarities for saliency detection. In *CVPR*, 478–485. IEEE.

Cong, Y.; Yuan, J.; and Liu, J. 2013. Abnormal event detection in crowded scenes using sparse representation. *Pattern Recognition* 46(7):1851–1864.

Dollár, P.; Rabaud, V.; Cottrell, G.; and Belongie, S. 2005. Behavior recognition via sparse spatio-temporal features. In *2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005*, 65–72. IEEE.

Hou, X., and Zhang, L. 2008. Dynamic visual attention: Searching for coding length increments. In *NIPS*, 681–688.

Kim, J., and Grauman, K. 2009. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *CVPR*, 2921–2928. IEEE.

Kratz, L., and Nishino, K. 2009. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models. In *CVPR*, 1446–1453. IEEE.

Li, W.; Mahadevan, V.; and Vasconcelos, N. 2014. Anomaly detection and localization in crowded scenes. *IEEE TPAMI* 36(1):18–32.

Lu, C.; Shi, J.; and Jia, J. 2013. Abnormal event detection at 150 fps in matlab. In *ICCV*, 2720–2727. IEEE.

Mahadevan, V.; Li, W.; Bhalodia, V.; and Vasconcelos, N. 2010. Anomaly detection in crowded scenes. In *CVPR*, 1975–1981. IEEE.

Mairal, J.; Bach, F.; Ponce, J.; and Sapiro, G. 2010. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research* 11:19–60.

Mehran, R.; Oyama, A.; and Shah, M. 2009. Abnormal crowd behavior detection using social force model. In *CVPR*, 935–942. IEEE.

Olshausen, B. A., and Field, D. J. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381(6583):607–609.

Pati, Y. C.; Rezaiifar, R.; and Krishnaprasad, P. 1993. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *IEEE 27th Annual Asilomar Conference on Signals, Systems, and Computers*, 40–44. IEEE.

Roshtkhari, M. J., and Levine, M. D. 2013. An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions. *Computer Vision and Image Understanding* 117(10):1436–1452.

Rubinstein, R.; Zibulevsky, M.; and Elad, M. 2008. Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit. *CS Technion* 40.

Saligrama, V., and Chen, Z. 2012. Video anomaly detection based on local statistical aggregates. In *CVPR*, 2112–2119. IEEE.

Wu, S.; Moore, B. E.; and Shah, M. 2010. Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes. In *CVPR*, 2054–2060. IEEE.

Zhao, B.; Fei-Fei, L.; and Xing, E. P. 2011. Online detection of unusual events in videos via dynamic sparse coding. In *CVPR*, 3313–3320. IEEE.