A Day on Campus - An Anomaly Detection Dataset for Events in a Single Camera

Mantini Pranav, Li Zhenggang, and Shah Shishir K

University of Houston, Houston TX 77004, USA
{pmantini, zli36}@uh.edu, shah@cs.uh.edu

Abstract. Detecting anomalies in videos is a complex problem with a myriad of applications in video surveillance. However, large and complex datasets that are representative of real-world deployment of surveillance cameras are unavailable. Anomalies in surveillance videos are not well defined and the standard and existing metrics for evaluation do not quantify the performance of algorithms accurately. We provide a large scale dataset, A Day on Campus (ADOC[1]), with 25 event types, spanning over 721 instances and occurring over a period of 24 hours. This is the largest dataset with localized bounding box annotations that is available to perform anomaly detection. We design a novel metric to evaluate the performance of methods and we perform an evaluation of the state-of-the-art methods to ascertain their readiness to transition into real-world surveillance scenarios.

1 Introduction

Surveillance cameras have become an integral part of public and private infrastructures. They provide a mechanism to actively monitor spaces for events and consequently enhance security. The advancement of sensor technology and the availability of affordable sensors has led to frequent deployment of large camera surveillance networks. Today, a university can have up to a thousand cameras, cities and casinos can have up to tens of thousands of cameras. These networks produce large quantities of video data and it is not possible to manually monitor and identify events. Computer vision algorithms that analyze videos find a natural place in these scenarios.

Given all events that may occur in the view of a surveillance camera, one would like to analyze the video to identify a subset of events that require attention. If the subset of events were known, the problem reduces to that of event detection [1, 2], where the goal is to model the events and identify them in videos. However, the events that may occur in a camera view are conditioned on a multitude of contextual factors, like view-point, geo-spatial factors, and etc. These factors are typically unknown and vary from camera to camera as well as events of interest.

[1] Dataset available at qil.uh.edu/datasets

To address this challenge researchers have resorted to the idea of identifying discordant observations. Edgeworth [3] described discordant observations "as those which present the appearance of differing in respect of their law of frequency from other observations with which they are combined" [3]. Most methods have defined anomaly as a deviation from the normal [4]. Given all observations that may occur and their frequency of occurrences, normal events can be defined as those that have a higher frequency of occurrence, and conversely **anomalous events** are defined as the complementary set of normal events (those that have lower frequency of occurrence). We will extrapolate and argue that the less probable an event is, the more anomalous it is. While we acknowledge that extremely less probable events do not necessarily imply that they are of greater significance from a video surveillance stand point.

1.1 Challenges in Anomaly Detection

Datasets for Anomaly Detection: Today vision algorithms are heavily data driven. One can see the difficulty in gathering a dataset of anomalous events. Such events are not well defined and are defined with respect to normal events, which again are not well defined. Furthermore, anomalous events have a low probability of occurrence and add to this difficulty. Figure 1 shows an example of an image acquired from a surveillance camera deployed at a university campus. An event can be anything happening in the scene including global events such as a weather anomaly to localized events such as a person falling down. We concern ourselves with events that are performed by humans, are affecting humans, and are consequences of human actions. The set of such events may not be independent of other events. For example, a person running can be conditioned on the fact that it is raining. However, we assume that the set of events concerning humans is independent.

Fig. 1. Anomalous Events.

With this assumption, we define events as holistic actions realized over time, and not their realization at individual time instants. For example, in the scenario shown in Figure 1 (left), the most common action is humans walking across. The blue bounding box shows a person distributing information (in the form of fliers) and their corresponding trajectory (shown in red). While the action of this person at any time instant can be looked as either standing or walking. The holistic action of distributing information is less probable compared to the action of a person walking, and is arguably more anomalous. Another example

in the image is the red bounding box that shows a crowd gathered around another person holding a sign. While the individual realization of the people in the crowd is the action of standing, the holistic action of crowd gathering as a consequence of a person holding a sign is unique. Most existing datasets do not contain such complex events, and label simple actions as anomalies. We distinguish them as separate events, where a person walking and standing could have a high frequency of occurrence, and a crowd gathering and a person distributing information are events of low frequency. It is in this respect that we distinguish our definition of anomalous events, and provide a dataset in pursuit of solutions for anomaly detection.

Anomaly Detection in Surveillance Cameras: Surveillance cameras are deployed in a variety of scenarios and are expected to function through varying global conditions like natural illumination changes such as day, night, dawn, etc. and weather changes such as rain, snow, and fog. They are deployed in a variety of scenarios such as indoor, outdoor, crowded areas, etc. Such events also result in a deviation from normal and often tend to produce false positives. A dataset for anomaly detection in surveillance videos should include sufficient variety of global changes, and low and high frequency events to test the effectiveness of algorithms.

Evaluation criteria: Existing evaluation metrics evaluate anomaly detection as a binary class (normal and anomaly) problem. The current evaluation schemes evaluate how well a method is capable of detecting anomalies while producing the least false positives. They tend to ignore the probability of occurrence of the event. Anomalies are not well defined, and the algorithms can encounter novel anomalies. We argue that it is advantageous to quantify the ability of the algorithm to detect anomalous events based on their probability of occurrence. For example, in the scenario shown in Figure 1, while it is not as probable as the action of a person walking, it is common to notice a person riding a bicycle. It is much less probable to notice a person walking a dog than a person riding a bicycle. We argue that an algorithm that detects the event that a person is walking a dog, and misses a few detections of a person riding a bicycle is more efficient than one that efficiently detects a person riding a bicycle and misses the less probable events. Most current evaluations weigh the detection of the person riding an bicycle and a person walking a dog equally. We propose an evaluation criteria to account for the probability of occurrence of anomalous events. In this paper we aim to address the above challenges. Our contributions are:

- We introduce a large surveillance dataset consisting of 24H of video from a single camera for anomaly detection with event annotations.
- We introduce an new evaluation criteria for anomaly detection algorithms.
- We perform benchmarking using state-of-the-art algorithms.

2 Existing Datasets

Anomaly detection in videos has been a widely researched problem in vision for decades. Traditional approaches were aimed at modeling local [10, 11, 7] and

Table 1. Comparison of existing dataset.

Datasets	# Abnormal	# Abnormal Events	# Scenes	Annotations	hours
UMN [5]	1,222	3	3	1222 Frame level	∼0.07
SubwayExit **[6]**	720	9	1	720 Frame level	∼0.72
SubwayEntrance **[6]**	2,400	21	1	2400 Frame level	1.6
Avenue **[7]**	3,820	47	1	3820 Bounding boxes	∼0.5
Ped1 **[4]**	4,005	40	1	2000 Pixel masks	1.5
Ped2 **[4]**	1,636	20	1	2100 Pixel masks	0.4
ShanghaiTech [8]	17,090	130	13	40791 Pixel masks	3.6
Streetscene [9]	19585	205	3	19585 Bounding boxes	4
ADOC	**97030**	**721**	1	**284125** Bounding boxes	**24**

holistic [12, 13] features to perform classification. More recently, various deep learning architectures have been used for anomaly detection viz. Convolutional Neural Networks (CNNs) [14], Long Short Term Memory Networks [15], Auto-encoders (AE), and Generative Adversarial networks (GANs) [16]. For a detailed review of the approaches for anomaly detection we direct readers to existing surveys [17–19].

Anomaly detection is used interchangeably to address two problems. First, a video classification problem, where given numerous videos, the task is to identify if a video contains anomaly or not. The latter is a temporal analysis problem, where, given continuous video from a camera, the task is to decide if each frame is either normal or anomalous. Anomaly detection is challenging because there is no consensus on what the exact definition is. Known activity labels including those to describe criminal or malicious activities is one definition. The other is one that represents rare events, or events that have a lower probability of occurrence. We use the latter in this paper, and hence, we refrain from using the term anomalous events in the paper and rather define them as low-frequency events. Sultani *et al.* [20] proposed a dataset for anomaly detection called UCF-Crime dataset that captures criminal activity well, while we focus on capturing the natural frequency of all human-related events in the scene of a surveillance video. Simple events such as "riding a bike" may not be a meaningful event under the first definition, but if we can construct algorithms that can understand the frequency of this event remains as an open problem. The UCF crime dataset has 128 hours from 1900 videos. The average video length is 4 minutes (70% of videos are $\leq$ 3 mins long), making it a valuable dataset for the task of video classification. However, this dataset is unsuitable for the task of temporal analysis. Furthermore, major reason vision algorithms fail to transition to real-world is that they produce too many false alarms. There is no comprehensive way of assessing the number of false positives an algorithm produces over a persisted amount of time (eg. a day) when deployed on a single camera. The proposed dataset is the only dataset that provides data for illumination and weather changes for a complex scene throughout an entire day. The proposed dataset is more in alignment with UMN [5], Subway [6], Avenue [7], etc.

Since our objective is to introduce a benchmark dataset that will spur research in anomaly detection as a comprehensive, multi-faceted problem. We

briefly review existing datasets and compare their size, scene, complexity and diversity. Table 1 shows the list of existing datasets.

In comparison, we introduce the largest dataset for anomaly detection in video surveillance cameras, consisting of 721 anomalous events localized using 284125 bounding box annotations. The dataset consists of 259123 frames captured over 24 contiguous hours and encapsulates the complexities of a typical surveillance camera.

3 A Day on Campus

The data is acquired from a surveillance camera deployed on a large university campus. It overlooks a walkway leading to various buildings and captures the events performed by students, faculty, and staff on a busy day. The camera captures video at a resolution of 1080p and a frame-rate of 3 frames per second. The video is compressed using H.264 format, which is a lossy compression method and is standard of the surveillance industry. We create a dataset from video captured over a period of 24 contiguous hours. The video encapsulates varying illumination conditions, crowded scenarios with background clutter. The data is annotated with a number of events ranging from low to high frequency.

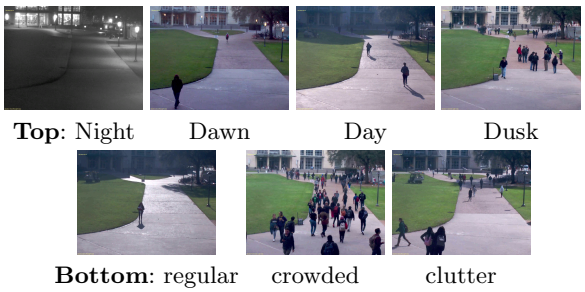


Fig. 2. Natural illumination changes occurring in camera view.

Data Variability: Outdoor surveillance cameras undergo regular variations in the scene throughout the day due to illumination changes. Figure 2 (Top) shows example images from the camera, representative of the captured changes in illumination through the day. The camera switches to infrared (IR) mode in extremely low illumination conditions which is typical of surveillance cameras, this produces a large shift in global image features. Currently, there are no existing datasets for anomaly detection that captures such variations in surveillance cameras. Furthermore, crowded scenarios, background clutter, and occlusion affect the performance of computer vision algorithms. Figure 2 (Bottom) shows example of the regular view of the camera (left), crowded (middle), and view with cluttered background (right: top right corner has a walkway and driveway with various people and cars moving by)

Events and Their Frequencies: There are a wide variety of events that occur in the scene from students walking, riding a bicycle or skateboard, staff driving

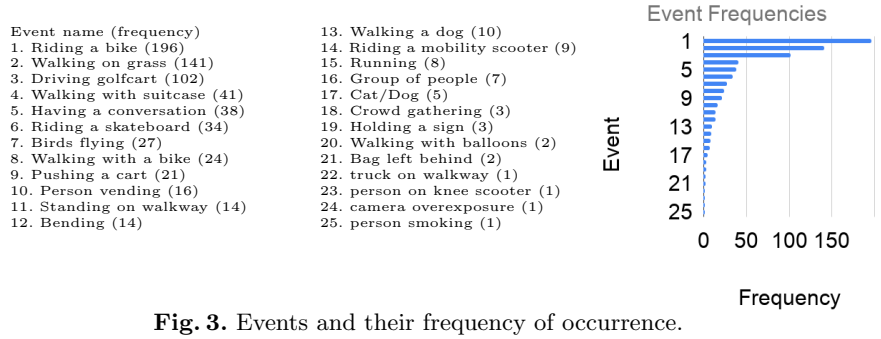


Fig. 3. Events and their frequency of occurrence.

golf carts to an occasional student walking a dog or carrying balloons. Figure 3 shows a list of the events and their distribution of occurrence rate. There are a total of 721 events in the scene that are considered as low frequency events. The most common activity in the scene is persons walking from one end to the other along the walkway. We estimate a total of 31675 people walking. This is considered as a high frequency event. If we consider the low frequency events as abnormal and the high frequency events as normal. Then the probability of an abnormal event is 0.022.

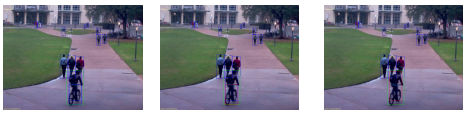
Event Annotations: The data is annotated using a combination of manual and automatic techniques.

Fig. 4. Manual Annotations.

Annotating Low Frequency Events: We manually annotate each low frequency event in the dataset. The annotations are performed using Computer Vision Annotation Tool (CVAT). We annotate key frames for each event and the annotation for the frames between are obtained through interpolation. Each event is annotated and then reviewed for accuracy by two annotators with computer vision background. Figure 4 shows the annotation for an event (riding a bicycle). The annotations are provided in the Multiple Object Tracking (MOT) format. Every event is localized using a tight bounding box, a frame id and contains a unique association ID to form trajectories.

Annotating High Frequency Events: We annotate the high frequency event of walking using automated technique. Given the set of all events occurred on the day, we assume that all the events excluding walking are low frequency and have already been annotated manually. We perform object detection and refine the results to automatically annotate walking. The results of the object detector are refined as shown in Figure 5 to generate automatic annotations. In step 1, any object that is not a person (car, bicycle, etc) is considered a low frequency

event and has been manually annotated. The output from the object detector is filtered and the bounding boxes labeled as person are retained and other are removed. In Figure 5, the unrefined images show all the detections from the object detector in blue. The green bounding box shows the manual annotation of the person riding a bike. The object detector detects the person on the bicycle, and the bicycle (among other detections). In the first step, all the bounding boxes that are not humans are removed. The resulting annotations are shown in the center image. The set of bounding boxes represents all events that involve



Unrefined step 1: Refined step 2:Refined

Fig. 5. Automatic Annotations. Blue BB: Automatic annotations, Green: Manual annotations, and Red: Deleted detections

humans including low and high frequency events. In the second step, we compute a disjoint set of bounding boxes from the manual annotations by removing all bounding boxes that have and Intersections Over Union (IOU) greater than a threshold value (0.2). The resulting set is assumed to be people walking. This is shown in Figure 5, the detection from object detector for the person on the bicycle is removed ($IOU \geq 0.2$). The deleted detections are shown in the red bounding box.

Fig. 6. Estimating frequency of people walking, yellow: bounding box is crossing the line, red: bounding box is not crossing the line

We rely on the accuracy of the object detector for these annotations. Anomaly detection algorithms designed for this scene would consider walking as normal events and is not aimed at detecting and localizing such occurrences, the accuracy of the detection are irrelevant. The annotations provide a way to estimate the comprehensive counts of each event, and subsequently estimate their probability of occurrences. Furthermore, unlike other datasets, they can enable the estimation of the false positives produced as a result of humans walking. The automatic annotations are provided in the YOLO format [21] and they do not contain any association across frames to form trajectories of walking. Figure 5 shows an example of automatically generated and filtered annotations.

Estimating the count of people walking: We estimate the total frequency of people walking to infer the probability of each event, and subsequently perform a more accurate evaluation of the performance of the algorithm. Most people that are annotated and that are detected are within this area of interest marked

by the red box in Figure 6 (left).To count the total number of people walking across this area. We first perform a perspective projection using the parallel lines in the scene to obtain a birds eye view as shown in Figure 6 (right) we project the area marked by the red box. Then, we estimate the count of total people walking over the entire video by counting, all the bounding boxes that cross the green line. Given the average pace of a pedestrian, we manually estimate the count of average number of frames over which a pedestrian crosses the line. We divide the total count by the number of bounding boxes detected crossing the line by this number to estimate the total number of walking events.

Numbers: Table 2 lists the count of frames, annotations, and events in the dataset. The dataset is made of 259123 frames, of which 97030 have low frequency events, and 142962 have either low, high or both frequency events. There are a total of 721 events annotated in the dataset. The 721 events are annotated using 13290 manually annotated bounding boxes, and 270835 interpolated bounding boxes. There are total of 284125 annotations representing low frequency events in the dataset. There are a total of 5082993 annotations that capture 31675 human walking over a period of 24 hours.

Table 2. Count of frames, annotations, and events in the dataset.

Description	Count
Low frequency events	
Trajectories	721
Manually Annotated bounding boxes	13290
Interpolated bounding boxes	270835
Total bounding boxes	284125
High frequency events	
Automatically annotated bounding boxes	5082993
Estimated number of persons walking	31675
Frame Count	
Frames with low frequency events	97030
Frames with events (low and high)	142962
Total Frames	259123

4 A Revised Evaluation Metric

Anomaly detection algorithms are evaluated as a binary class problem, where anomalous events are considered as the positive class and the normal as negative. By our definition, anomalous events occur with a low probability. In general one can assume that the distribution is biased towards the negative class. For example, in the ADOC dataset, there are 31675 samples that belong to negative class and 721 samples that belong to positive class. Currently accuracy and precision are the common metrics to evaluate the performance of the algorithms. Accuracy is defined as $\frac{TP+TN}{TP+FP+TN+FN}$, where TP are true positives, FP are false positives, TN are true negatives, and FN are false negatives. An algorithms that detects all events as belonging to the negative class and fails to detect any anomalous event has an accuracy of 0.97 $((0 + 31675)/32396)$. Precision is defined as $\frac{TP}{TP+FP}$, While this metric can quantify the ability of the system to detect anomalies, it does not capture the ability of the system to identify negative samples. It is necessary to quantify the overall capability of the system by aggregating the algorithms capability to detect individual types of events.

Let there be n events that may occur in a surveillance scenario, denoted by $\{e_1, e_2, ..., e_n\}$ with probabilities $\{p_1, p_2, ..., p_n\}$ such that $\Sigma_i p_i = 1$. Now given various algorithms, according to Expected Utility Theory, the decision of which algorithm to choose is dictated by maximizing the expected utility [22].

$$Eu(A) = \Sigma_i p_i U(e_i), \tag{1}$$

where $Eu(A)$ is the expected utility for the algorithm, and $U(e_i)$ is an assigned numerical utility.

For example, if $\{a_1, a_2, ..., a_n\}$ be the accuracy with which the algorithm A detects events $\{e_1, e_2, ..., e_n\}$, respectively. If $U(e_i) = a_i$, then the expected utility reduces to that of computing the expected accuracy $E_a = \Sigma_i p_i a_i$.

Note that similar to computing the accuracy under a binary class assumption, the expected accuracy is high when the system is capable of detecting the high probability events, and the effect of detecting low probability events with a high accuracy is insignificant.

Our motivation to quantify the ability of an algorithm to perform anomalous event detection has roots in prospect theory [23], where we depart from the expected utility theory by overweighting the small probabilities and underweighting the large [24]. The decision of one algorithm over the other is defined by the prospect value $\pi_A = \Sigma_i w_\gamma(p_i) a_i$, where $w_\gamma(p)$ is the probability weighting function. A simple weighting function is to assume that all events occur with equal probability, i.e. $w_\gamma(p_i) = 1/n$, where the prospect value reduces to the average of the accuracy of the algorithms ability to detect each event. However, considering that the normal class can have a large number of samples, and such probability weighting can undermine the effect of the algorithms inability to detect these events. This can translate to choosing algorithms that produce too many false positives.

Linear in Log Odds (lilo): We adopt the two parameter probability weighting function defined as [25]:

$$w_\gamma(p) = \frac{\delta p^\gamma}{(\delta p^\gamma + (1-p)^\gamma)}, \tag{2}$$

where the γ controls the curvature and δ controls the elevation independently. Note when $\delta = 1$, and $\gamma = 1$ the function reduces to $w_\gamma(p) = p$. Figure 7 show the weighting function for $\delta = 1$, and $\gamma = 0.2$. We compute the expected accuracy value using lilo weighted probabilities (liloAcc) as a metric to quantify the performance of anomaly detection methods. Figure 7 (right) shows the transformed probabilities using the lilo function for each event in the dataset.

Note that the liloAcc metric is bounded, as the accuracy and the probability values are bounded. The maximum value occurs when the accuracy of detecting each event is 1. The maximum value is $\Sigma_i w_\gamma(p_i)$, sum of the lilo weighted probabilities. We use this value to obtain a normalized metric (nliloAcc):

$$nliloAcc = \frac{\Sigma_i w_\gamma(p_i) a_i}{\Sigma_i w_\gamma(p_i)} \tag{3}$$

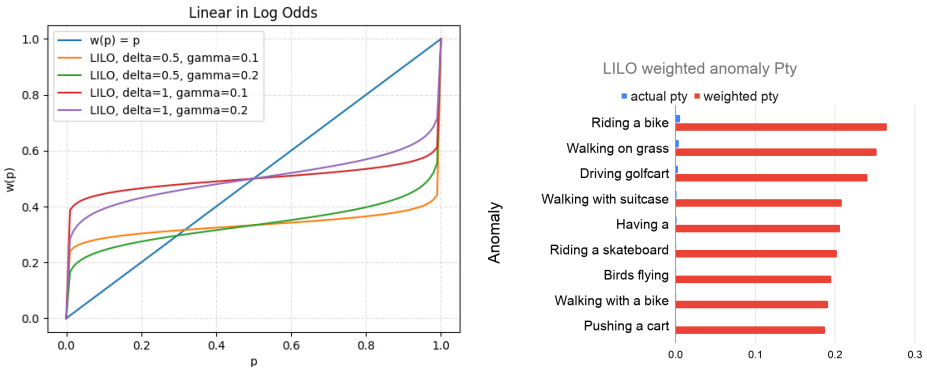


Fig. 7. (left) Linear in Log Odds function, (right), few event probabilities in ADOC dataset weighted by lilo (blue: original probability values, red: lilo weighted probability values)

5 Experimental design

The goal of the experiments is to ascertain the performance of existing state-of-the-art video anomaly detection algorithms on ADOC dataset. The dataset includes a variety of complex events that are novel, and numerous illumination changes. The experiments provide insight into the readiness of the existing algorithms to transition to real time analysis for anomaly detection.

The dataset consists of 259123 frames captured over 24 hours of video. We divide the dataset into two 12 hour parts and use them as training and testing patterns. Both partitions contain a combination of day time and night time images. To account for this complexity, we separate the dataset into two partitions, first containing exclusively day images and the latter night. We quantify the capability of algorithms to perform in day time, night time, and overall. We perform the following experiments:

- Experiment **1**: Training: **Day**, Testing: **Day** images
- Experiment **2**: Training: **Night**, Testing: **Night** images
- Experiment **3**: Training: **Day+Night**, Testing: **Day+Night** images

5.1 State of the art methods

We choose popular methods from the recent literature and evaluate their performance on the ADOC dataset. We choose methods that have been largely cited and ascertain their readiness to transition into real world analysis of surveillance videos. Anomaly detection methods can be designed to capture a variety of features. Some methods can exclusively encode spatial information or they can encode a combination of spatial and temporal information. Some methods perform analysis on image patches and accumulate information to make decisions while others make an inference on the entire frame. We choose one method that performs analysis on patches by encoding only spatial information, and two other methods that make frame level decisions and encode temporal patterns for anomaly detection. We evaluate the following methods:

- Abnormal Event Detection in Videos using Spatiotemporal Autoencoder (SPAE) (ISNN 2017): The method consists of a spatial feature extractor and a temporal encoder-decoder, which together learn the temporal patterns in videos. This method processes images as a whole and captures temporal patterns among them. We use an input frame size of 227X227 for training and testing.
- Adversarially Learned One-Class Classifier for Novelty Detection [26] (ALOCC) (CVPR 2018): This work proposes a method for novelty detection by training a one class classifier. The framework consists of two modules. The former acts as a pre-processing step for representation learning. The latter performs discrimination to detect novel classes. The paper uses two methods to perform anomaly detection, where a test image is input to the discriminator and the likelihood value is used to detect anomalies. In the second method, the likelihood value for the reconstructed method is used to detect anomaly. Results from the experiments performed in this paper demonstrate that the latter outperforms the first. If D represents the discriminator and R the encoder-decoder. In the first, the value $D(X)$ is compared against a threshold to infer if X is an anomalies. In the second the value $D(R(X))$ is thresholded to make a decision. We evaluate both methods. The input to the network are patches of images. We use an input frame size of 256X354 for training and testing. We set the patch size to 64X64.
- Future Frame Prediction for Anomaly Detection – A New Baseline [8] (FFP) (CVPR 2018): This method proposes to predict frames into the future, and exploit the inability to predict anomalous events to detect them. A good prediction would imply that it is normal, and vice versa. A UNET is trained by minimizing a loss function that encodes both spatial intensity features and temporal features (optical flow). This method performs analysis at a temporal level and encode both temporal and spatial information for anomaly detection. We use an input frame size of 256X256 for training and testing.
- Memorizing Normality to Detect Anomaly [28] (MemAE) (ICCV 2019): The overall approach involves augmenting an autoencoder with a memory module that is records the prototypical elements of the encoded normal data. This method consist of an encoder, a decoder, and a memory module. The input image is encoded, which is used to retrieve the relevant items from the memory module, which is then reconstructed by the encoder. A larger reconstruction error implies an anomaly. We use the the pre-trained weights on the Ped2 [4] dataset available from the authors for this evaluation.
- Learning Memory-guided Normality for Anomaly Detection [29] (Mnad) (CVPR 2020): This method is similar to the MemAE [28]. It improves the memory module by recording diverse and discriminative normal patterns, by separating memory items explicitly using feature compactness and separateness losses, and enabling using a small number of items compared to MemAE (10 vs 2,000 for MemAE). We use the the pre-trained weights on the Ped2 [4] dataset available from the authors for this evaluation.

5.2 Benchmarks

We compare and quantify the performance of the algorithms under a two class assumption, and then compare them using the proposed lilo weighted expected accuracy metric. The decision of normal and abnormal is made at each level or for a series of frames. Assuming the normal frames to be the negative class, and abnormal as the positive, we quantify the overall accuracy by plotting the Receiver Operating Characteristic (ROC) curves along with their corresponding area under the curve (AUC) as shown in Figure 8. Then we quantify the capability of each method at the optimal threshold. We define optimal threshold as the point on the ROC curve where the difference between TPR and FPR is maximum. The threshold at which the method detects maximum anomalies while reducing false positives. The results are shown in Table 3.

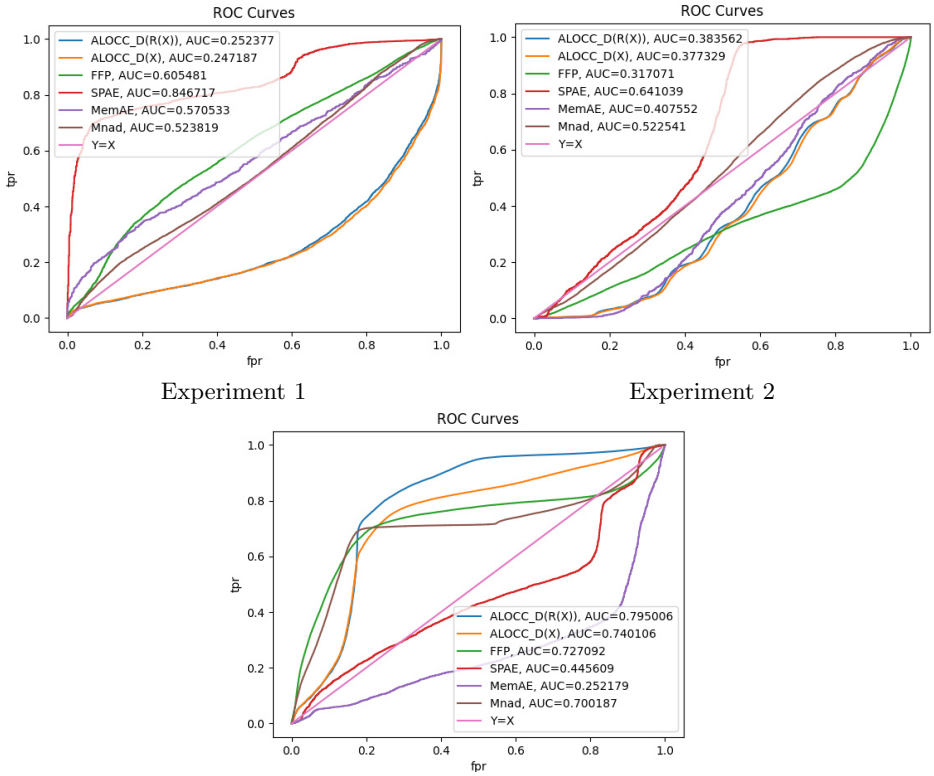


Fig. 8. ROC Curve

Experiment 1: Three models are trained and tested on exclusively day time images from the dataset. ALOCC produces a large number of false negatives. FFP does better at prediction than ALOCC, but produces too many false positives. SPAE captures the largest number of true positives, compared to other methods.

SPAE has an AUC score of 0.846, with an accuracy of 0.735 and outperforms the other models. Figure 9 shows a bar plot where the x-axis consists of various events ordered by their occurrence rate in ascending order. The plot shows the accuracy of each method in detecting a particular event. SPAE shows an ability to detect events of varying frequency and this is reflected in the nlilo Accuracy score. SPAE and FFP perform temporal analysis while ALOCC performs only spatial analysis. This suggests that temporal analysis is essential when detecting anomalous events.

Experiment 2: Three models are trained and tested on exclusively night images. ALOCC and SPAE tend to produce a large number of false positives. FFP labels most of all frames as negative, and fails to detect abnormal events. SPAE produces a significantly lower false positives than ALOCC. The AUC score suggests that SPAE performs best in this scenario. Despite the fact that FFP fails to detect abnormal events, the accuracy score suggests that it performs best followed by SPAE. Figure 9 shows that SPAE captures abnormal events across varying frequencies best, followed by ALOCC and the FPP. The nlilo Accuracy suggests that SPAE is the best performing algorithm with its relatively high true positives and significantly low false positives. This demonstrates the need for probability distorted metric to evaluate anomaly detection methods.

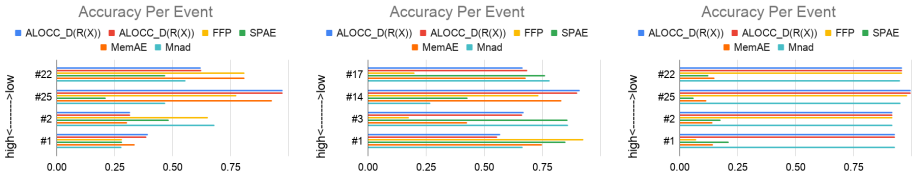
Experiment 3: In this experiment three models are trained and tested on

Table 3. Comparison of performance the state-of-the-art methods at optimal threshold, Acc - accuracy, liloAcc - lilo weighted expected accuracy.

Experiment	Method	Year/conf	TN	FN	TP	FP	AUC	Acc	nliloAcc
1	$ALOCC_{D(R(X))}$	CVPR 18'	10432	53072	1083	113	0.257	0.178	0.299
	$ALOCC_{D(X)}$	CVPR 18'	10427	52970	1185	118	0.247	0.179	0.307
	FFP	CVPR 18'	7477	29166	24989	3068	0.605	0.501	0.583
	$SPAE$	ISNN 17'	9831	16429	37726	714	0.846	0.735	0.698
	$MemAE$	ICCV 19'	8381	35283	18872	2164	0.570	0.421	0.598
	$Mnad$	CVPR 20'	9082	43607	10548	1463	0.523	0.303	0.588
2	$ALOCC_{D(R(X))}$	CVPR 18'	1187	191	20581	48441	0.383	0.309	0.627
	$ALOCC_{D(X)}$	CVPR 18'	1352	202	20570	48276	0.377	0.311	0.656
	FFP	CVPR 18'	49625	20766	6	3	0.346	0.704	0.583
	$SPAE$	ISNN 17'	22226	514	20258	27402	0.641	0.603	0.819
	$MemAE$	ICCV 19'	3080	608	20164	46548	0.407	0.330	0.596
	$Mnad$	CVPR 20'	12041	3239	17533	37587	0.522	0.420	0.651
3	$ALOCC_{D(R(X))}$	CVPR 18'	45071	14845	60082	15102	0.795	0.778	0.824
	$ALOCC_{D(X)}$	CVPR 18'	44415	19042	55885	15758	0.740	0.742	0.795
	FFP	CVPR 18'	47733	22835	52092	12440	0.727	0.738	0.824
	$SPAE$	ISNN 17'	53374	63426	11501	6799	0.445	0.480	0.309
	$MemAE$	ICCV 19'	60173	74911	16	0	0.252	0.445	0.192
	$Mnad$	CVPR 20'	49589	23294	51633	10584	0.700	0.749	0.881

a combination of day time and night time images together. All methods produce higher false positives compared to Experiment 1. MemAE labels most of all frames as negative, and fails to detect abnormal events. AUC and Accuracy scores suggests ALOCC perform better than other methods. Interestingly, Unlike the first two experiments ALOCC outperforms SPAE in Experiment 3. Our inference is that the method generalizes better on exposing to varied illumination. Figure 9 shows that it is able to capture events across frequencies better than SPAE and is comparable to ALOCC. nlilo Accuracy aggregates a similar score for ALOCC and FPP. SPAE underperforms in Experiment 3. This suggests

that the model is inefficient at learning features and performing abnormality detection across illuminations. Furthermore, performing patch level analysis seems to be beneficial when learning representation across illuminations. It is able to capture local features better compared to other methods. We conclude that there is need to perform both temporal analysis and a patch level analysis to build robust algorithms for anomaly detection. Overall Mnad outperfroms the other methods in this experiment.



Experiment 1 Experiment 2 Experiment 3

Fig. 9. Bar plot showing the accuracy of detecting example individual events for each method in experiment 1,2, and 3, x-axis shows event numbers as used in Figure 3

Discussion: MemAE and Mnad use pretrained weights and are not trained on the adoc dataset (due to the lack of trianing code and optimal hyper-parameters) unlike the former four methods. Specifically training them on the adoc dataset can improve the detection accuracy. Experiment 1 demonstrates that the state-of-the-art methods fail at detection anomalies and tend to generate too many false negatives. Experiment 2 suggests that the methods tend to produce large numbers of false positives at night time. Experiment 3 suggests that there is much work needed towards designing algorithms that can adapt to varying illuminations. These improvements and the pursuit of robust methods designed with the goal of detecting holistic anomalous events is necessary to realize anomaly detection in surveillance cameras. Experiments also showcase the inconsistencies in adopting generic sensitivity and specificity metrics that are used for binary classification. Given the biased nature of anomaly detection datasets, we advocate the need for probability distorted metrics.

6 Conclusion

We have introduced a complex and a large scale dataset with video from a surveillance camera to enable development of robust algorithms for anomaly detection. We have defined events that are relevant to surveillance cameras in pedestrian environments, and provided a dataset with numerous event annotations. We have defined metrics to account for the biased nature of anomaly detection datasets. We have evaluated the state-of-the-art methods available for anomaly detection on the ADOC dataset. We have accumulated the results and established the required research direction to enable robust anomaly detection in surveillance videos.

References

1. Wang, F., Jiang, Y.G., Ngo, C.W.: Video event detection using motion relativity and visual relatedness. In: Proceedings of the 16th ACM international conference on Multimedia. (2008) 239–248
2. Zhang, D., Chang, S.F.: Event detection in baseball video using superimposed caption recognition. In: Proceedings of the tenth ACM international conference on Multimedia. (2002) 315–318
3. Edgeworth, F.Y.: Xli. on discordant observations. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **23** (1887) 364–375
4. Mahadevan, V., Li, W., Bhalodia, V., Vasconcelos, N.: Anomaly detection in crowded scenes. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE (2010) 1975–1981
5. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2009) 935–942
6. Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D.: Robust real-time unusual event detection using multiple fixed-location monitors. IEEE transactions on pattern analysis and machine intelligence **30** (2008) 555–560
7. Lu, C., Shi, J., Jia, J.: Abnormal event detection at 150 fps in matlab. In: Proceedings of the IEEE international conference on computer vision. (2013) 2720–2727
8. Liu, W., Luo, W., Lian, D., Gao, S.: Future frame prediction for anomaly detection–a new baseline. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6536–6545
9. Ramachandra, B., Jones, M.: Street scene: A new dataset and evaluation protocol for video anomaly detection. In: The IEEE Winter Conference on Applications of Computer Vision. (2020) 2569–2578
10. Saligrama, V., Konrad, J., Jodoin, P.M.: Video anomaly identification. IEEE Signal Processing Magazine **27** (2010) 18–33
11. Bertini, M., Del Bimbo, A., Seidenari, L.: Multi-scale and real-time non-parametric approach for anomaly detection and localization. Computer Vision and Image Understanding **116** (2012) 320–329
12. Marsden, M., McGuinness, K., Little, S., O'Connor, N.E.: Holistic features for real-time crowd behaviour anomaly detection. In: 2016 IEEE International Conference on Image Processing (ICIP), IEEE (2016) 918–922
13. Xie, S., Guan, Y.: Motion instability based unsupervised online abnormal behaviors detection. Multimedia Tools and Applications **75** (2016) 7423–7444
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9** (1997) 1735–1780
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. (2014) 2672–2680
17. Pawar, K., Attar, V.: Deep learning approaches for video-based anomalous activity detection. World Wide Web **22** (2019) 571–601
18. Chalapathy, R., Chawla, S.: Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407 (2019)

19. Kumaran, S.K., Dogra, D.P., Roy, P.P.: Anomaly detection in road traffic using visual surveillance: A survey. arXiv preprint arXiv:1901.08292 (2019)
20. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6479–6488
21. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv (2018)
22. Bernoulli, D.: Exposition of a new theory on the measurement of risk. In: The Kelly capital growth investment criterion: Theory and practice. World Scientific (2011) 11–24
23. Kahneman, D., Tversky, A.: Prospect theory: An analysis of decision under risk. In: Handbook of the fundamentals of financial decision making: Part I. World Scientific (2013) 99–127
24. Glaser, C., Trommershäuser, J., Mamassian, P., Maloney, L.T.: Comparison of the distortion of probability information in decision under risk and an equivalent visual task. Psychological science **23** (2012) 419–426
25. Gonzalez, R., Wu, G.: On the shape of the probability weighting function. Cognitive psychology **38** (1999) 129–166
26. Sabokrou, M., Khalooei, M., Fathy, M., Adeli, E.: Adversarially learned one-class classifier for novelty detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 3379–3388
27. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, Springer (2015) 234–241
28. Gong, D., Liu, L., Le, V., Saha, B., Mansour, M.R., Venkatesh, S., Hengel, A.v.d.: Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 1705–1714
29. Park, H., Noh, J., Ham, B.: Learning memory-guided normality for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 14372–14381