Continual Learning for Anomaly Detection in Surveillance Videos

Keval Doshi
University of South Florida
4202 E Fowler Ave, Tampa, FL 33620
kevaldoshi@mail.usf.edu

Yasin Yilmaz
University of South Florida
4202 E Fowler Ave, Tampa, FL 33620
yasiny@usf.edu

Abstract

Anomaly detection in surveillance videos has been recently gaining attention. A challenging aspect of high-dimensional applications such as video surveillance is continual learning. While current state-of-the-art deep learning approaches perform well on existing public datasets, they fail to work in a continual learning framework due to computational and storage issues. Furthermore, online decision making is an important but mostly neglected factor in this domain. Motivated by these research gaps, we propose an online anomaly detection method for surveillance videos using transfer learning and continual learning, which in turn significantly reduces the training complexity and provides a mechanism for continually learning from recent data without suffering from catastrophic forgetting. Our proposed algorithm leverages the feature extraction power of neural network-based models for transfer learning, and the continual learning capability of statistical detection methods.

1. Introduction

The number of closed-circuit television (CCTV) surveillance cameras are estimated to go beyond 1 billion globally by the end of 2021 [22]. Particularly, video surveillance is an essential tool with applications in law enforcement, transportation, environmental monitoring, etc. For example, it has become an inseparable part of crime deterrence and investigation, traffic violation detection, and traffic management. However, the monitoring ability of surveillance systems has been unable to keep pace due to the massive volume of streaming video data generated in real-time. This has resulted in a glaring deficiency in the adequate utilization of available surveillance infrastructure and hence there is a pressing need for developing intelligent computer vision algorithms for automatic video anomaly detection.

Video anomaly detection plays an important role in ensuring safety, security and sometimes prevention of potential catastrophes, hence another critical aspect of a video anomaly detection system is the real-time decision making capability. Events such as traffic accidents, robbery, and fire in remote places require immediate counteractions to be taken promptly, which can be facilitated by the real-time detection of anomalous events. However, online and real-time detection methods have only recently gained interest [30]. Also, many methods that claim to be online heavily depend on batch processing of long video segments. For example, [23, 16] perform a normalization step which requires the entire video. Regarding the importance of timely detection in video, as [30] argues, the methods should also be evaluated in terms of the average detection delay, in addition to the commonly used metrics such as true positive rate, false positive rate, and area-under-the-curve (AUC).

Although deep neural networks provide superior performance on various machine learning and computer vision tasks, such as object detection [8], image classification [21], playing games [38], image synthesis[35], etc., where sufficiently large and inclusive data sets are available to train on, there is also a significant debate on their shortcomings in terms of interpretability, analyzability, and reliability of their decisions [17]. Recently, statistical and nearest neighbor-based methods are gaining popularity due to their appealing characteristics such as being amenable to performance analysis, computational efficiency, and robustness [4, 12].

A key challenge of anomaly detection in videos is that defining notions of normality and abnormality that encompass all possible nominal and anomalous data patterns are nearly impossible. Thus, for a video anomaly detection framework to work in a practical setting, it is extremely crucial that it is capable of learning *continually* from *a small number of new samples* in an online fashion. However, a vast majority of existing video anomaly detection methods are completely dependent on data-hungry deep neural networks [40]. It is well known that naive incremental strategies for continual learning in deep/shallow neural networks suffer from catastrophic forgetting [19]. On the other hand, a cumulative approach would require all previous data to be stored and the model to be retrained on the entire data.

This approach quickly becomes infeasible due to computational and storage issues. Thus, preserving previously learned knowledge without re-accessing previous data remains particularly challenging [25]. Recent advances in transfer learning have shown that using previously learned knowledge on similar tasks can be useful for solving new ones [24]. Hence, we propose a hybrid use of transfer learning via neural networks and statistical k-nearest neighbor (kNN) decision approach for finding video anomalies with limited training in an online fashion. In summary, our contributions in this paper are as follows:

- We leverage *transfer learning* to significantly reduce the training complexity while simultaneously outperforming current state-of-the-art algorithms.
- We propose a statistical framework for sequential anomaly detection which is capable of *continual* and *few-shot* learning from videos.
- We extensively evaluate our proposed framework on publicly available video anomaly detection datasets and also on a real surveillance camera feed.

In Section 2, we review the related literature for anomaly detection in surveillance videos. Section 3 describes the proposed method, a novel hybrid framework based on neural networks and statistical detection. In Section 4, the proposed method is compared in detail with the current state-of-the-art algorithms. Finally, in Section 5 some conclusions are drawn, and future research directions are discussed.

2. Related Works

A commonly adopted learning technique due to the inherent limitations in the availability of annotated and anomalous instances is semi-supervised anomaly detection, which deals with learning a notion of normality from nominal training videos. Any significant deviation from the learned nominal distribution is then classified as anomalous [5, 16]. On the other hand, supervised detection methods which train on both nominal and anomalous videos have limited application as obtaining the annotations for training is difficult and laborious. To this end, [40] proposes using a deep multiple instance learning (MIL) approach to train on video-level annotated videos, in a weakly supervised manner. Even though training on anomalous videos might enhance the detection capability on similar anomalous events, supervised methods would typically suffer in a realistic setup from unknown/novel anomaly types.

A key component of computer vision problems is the extraction of meaningful features. In video surveillance, the extracted features should capture the difference between the nominal and anomalous events within a video. The selection of features significantly impacts the identifiability of types of anomalous events in video sequences. Early techniques primarily focused on trajectory features [1], limiting their applicability to detection of anomalies related to moving objects and trajectory patterns. For example, [11] studied detection of abnormal vehicle trajectories such as illegal U-turn. [31] extracts human skeleton trajectory patterns, and hence is limited to only the detection of abnormalities in human behavior.

Another class of widely used features in this domain are motion and appearance features. Traditional methods extract the motion direction and magnitude to detect spatiotemporal anomalies [37]. Histogram of optical flow [3, 6], and histogram of oriented gradients [9] are some other commonly used hand-crafted feature extraction techniques frequently used in the literature. The recent literature is dominated by the neural network-based methods [13, 14, 23, 28, 33, 36, 44] due to their superior performance [44]. Contrary to the hand-crafted feature extraction, neural network-based feature extraction methods [44] learn the appearance and motion features by deep neural networks. In [27], the author utilizes a Convolutional Neural Networks (CNN), and Convolutional Long Short Term Memory (CLSTM) to efficiently learn appearance and motion features, respectively. More recently, Generative Adversarial Networks (GAN) have been gaining popularity as they are able to generate internal scene representations based on a given frame and its optical flow.

However, there has been a significant ongoing debate of the shortcomings of neural network-based methods in terms of interpretability, analyzability, and reliability of their decisions [17]. Furthermore, it is well known that neural networks are notoriously difficult to train on new data or when few samples of a new class are available, i.e., they struggle with continual learning and few-shot learning. Hence, recently few-shot learning and continual learning have been studied in the computer vision literature [20, 42, 39, 43, 25]. However, not a lot of progress has been made yet in the field of continual learning with applications to video surveillance. Hence, in this work, we primarily compare our continual learning performance with the state-of-the-art video anomaly detection algorithms even though they are not tailored for continual learning.

3. Proposed Method

3.1. Motivation

In existing anomaly detection in surveillance videos literature, an anomaly is construed as an unusual event which does not conform to the learned nominal patterns. However, for practical implementations, it is unrealistic to assume the availability of training data which takes all possible nominal patterns/events into account. Often, anomalous events are circumstantial in nature and it is challenging

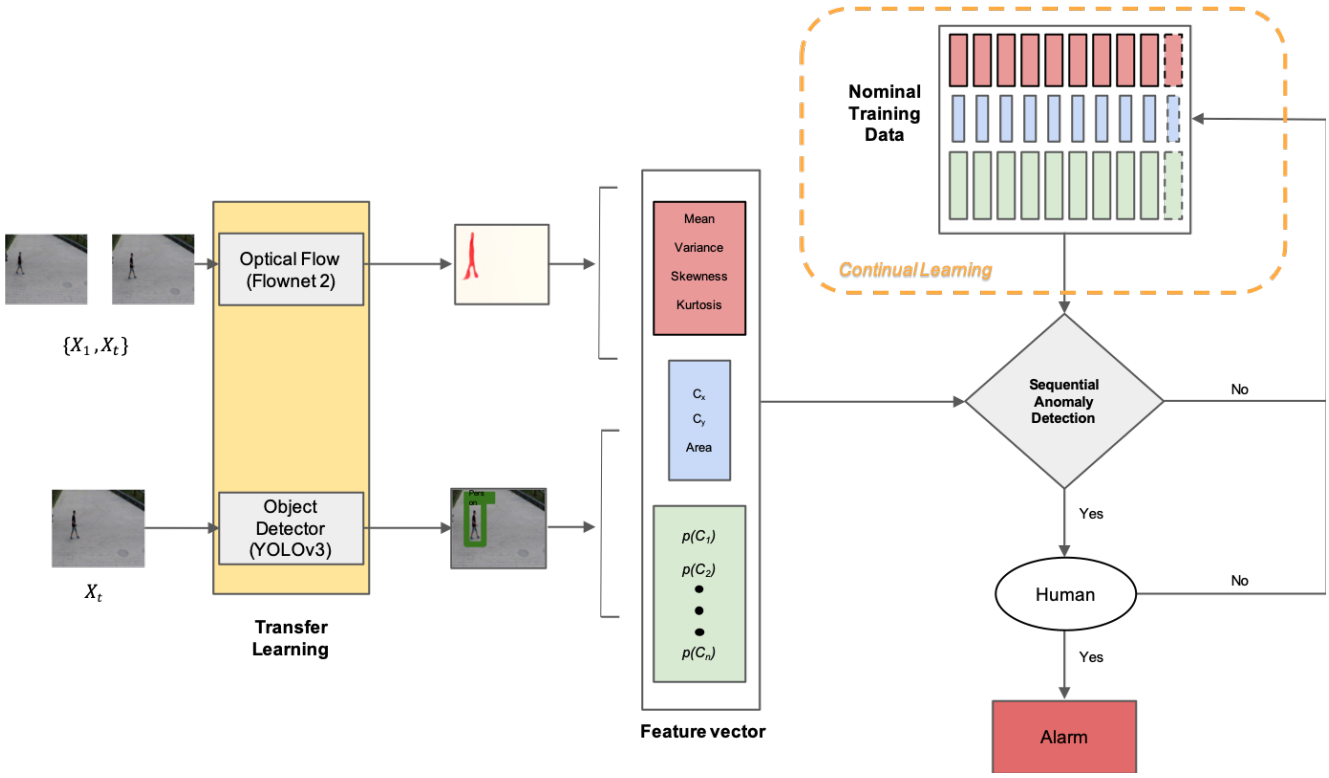

Figure 1. Proposed continual learning framework. At each time t, neural network-based feature extraction module provides motion (optical flow), location (center coordinates and area of bounding box), and appearance (class probabilities) features to the statistical anomaly detection module, which makes online decisions and continual updates to its decision rule.

to distinguish them from nominal events. For example, in many publicly available datasets a previously unseen event such as a person riding a bike is considered as anomalous, yet under different conditions, the same event can be categorized as nominal. Thus, a practical framework should be able to update its definition of nominal events *continually*. This presents a novel challenge to the current approaches mentioned in Section 2, as their decision mechanism is extensively dependent on Deep Neural Networks (DNNs). DNNs typically require the entire training data to be made available prior to the learning task as updating the model on new data necessitates either retraining from scratch , which is computationally expensive, or iteratively with the risk of catastrophic forgetting [19]. Moreover, another motivational fact for us is that the sequential nature of video anomaly detection and the importance of online decision making are not well addressed [30].

3.2. Feature Selection

Most existing works focus on a certain aspect of the video such as optical flow, gradient loss or intensity loss. This in turn restrains the existing algorithms to a certain form of anomalous event which is manifested in the considered video aspect. However, in general, the type of anomaly is broad and unknown while training the algorithm. For example, an anomalous event can be justified on the basis of appearance (a person carrying a gun), motion (two people fighting) or location (a person walking on the roadway). To account for all such cases, we create a feature vector F_t^i for each object i in frame X_t at time t, where F_t^i is given by $[w_1 F_{motion}, w_2 F_{location}, w_3 F_{appearance}]$. The weights w_1, w_2, w_3 are used to adjust the relative importance of each feature category.

3.3. Transfer Learning

Most existing works propose training specialized data-hungry deep learning models from scratch, however this bounds their applicability to the cases where abundant data is available. Also, the training time required for such models grows exponentially with the size of training data, making them impractical to be deployed in scenarios where the model needs to continually learn. Hence, we propose to leverage transfer learning to extract meaningful features from video.

Object Detection: To obtain location and appearance features, we use a pre-trained object detection system such as You Only Look Once (YOLO) [34] to detect objects in video streams in real time. As compared to other state-of-

the-art models such as SSD and ResNet, YOLO offers a higher frames-per-second (fps) processing while providing better accuracy. For online anomaly detection, speed is a critical factor, and hence we currently prefer YOLOv3 in our implementations. We get a bounding box (location), along with the class probabilities (appearance) for each object detected in frame X_t. Instead of simply using the entire bounding box, we monitor the center of the box and its area to obtain the location features. In a test video, objects diverging from the nominal paths and/or belonging to previously unseen classes will help us detect anomalies, as explained in Section 3.5.

Optical Flow: Apart from spatial information, temporal information is also a critical aspect of videos. Hence, we propose to monitor the contextual motion of different objects in a frame using a pre-trained optical flow model such as Flownet 2 [15]. We hypothesize that any kind of motion anomaly would alter the probability distribution of the optical flow for the frame. Hence, we extract the mean, variance, and the higher order statistics skewness and kurtosis, which represent asymmetry and sharpness of the probability distribution, respectively.

3.4. Feature Vector

Combining the motion, location, and appearance features, for each object i detected in frame X_t, we construct the feature vector

$$F_t^i = \begin{bmatrix} w_1\text{Mean} \\ w_1\text{Variance} \\ w_1\text{Skewness} \\ w_1\text{Kurtosis} \\ w_2\text{C}_\text{x} \\ w_2\text{C}_\text{y} \\ w_2\text{Area} \\ w_3 p(C_1) \\ w_3 p(C_2) \\ \vdots \\ w_3 p(C_n) \end{bmatrix}, \tag{1}$$

as shown in Fig. 1, where Mean, Variance, Skewness and Kurtosis are extracted from the optical flow; $\text{C}_\text{x}, \text{C}_\text{y}$, Area denote the coordinates of the center of the bounding box and the area of the bounding box from the object detector; and $p(C_1), \ldots, p(C_n)$ are the class probabilities for the detected object. Hence, at any given time t, with n denoting the number of possible classes, the dimensionality of the feature vector is given by $m = n + 7$.

3.5. Anomaly Detection

We aim to detect anomalies in streaming videos with minimal detection delays while satisfying a desired false alarm rate. Specifically for video surveillance, we can safely hypothesize that any anomalous event would persist for an unknown period of time. This makes the problem suitable for a sequential anomaly detection framework [2]. However, since we have no prior knowledge about the anomalous event that might occur in a video, traditional parametric algorithms which require probabilistic models and data for both nominal and anomalous cases cannot be used directly. Thus, we propose the following nonparametric sequential anomaly detection algorithm.

Training: Given a set of N training videos $\mathcal{V} \triangleq \{v_i : i = 1, 2, \ldots, N\}$ consisting of P frames in total, we leverage the deep learning module of our proposed detector to extract M feature vectors $\mathcal{F}^M = \{F^i\}$ for M detected objects in total such that $M \geq P$. We assume that the training data does not include any anomalies. These M vectors correspond to M points in the nominal data space, distributed according to an unknown complex probability distribution. Our goal here is to learn a nonparametric description of the nominal data distribution. We propose to use the Euclidean k nearest neighbor (kNN) distance, which captures the local interactions between nominal data points, to figure out a nominal data pattern due to its attractive traits, such as analyzability, interpretability, and computational efficiency [4, 12]. We hypothesize that given the informativeness of extracted motion, location, and appearance features, anomalous instances are expected to lie further away from the nominal manifold defined by $\mathcal{F}^M$. That is, the kNN distance of anomalous instances with respect to the nominal data points in $\mathcal{F}^M$ will be statistically higher as compared to the kNN distances of nominal data points. The training procedure of our detector is given as follows:

1. Randomly partition the nominal dataset $\mathcal{F}^M$ into two sets $\mathcal{F}^{M_1}$ and $\mathcal{F}^{M_2}$ such that $M = M_1 + M_2$.
2. Then, for each point F^i in $\mathcal{F}^{M_1}$, we compute the kNN distance d_i with respect to the points in set $\mathcal{F}^{M_2}$.
3. For a significance level α, e.g., 0.05, the $(1-\alpha)$th percentile d_α of kNN distances $\{d_1, \ldots, d_{M_1}\}$ is used as a baseline statistic for computing the anomaly evidence of test instances.

Testing: During the testing phase, for each object i detected at time t, the deep learning module constructs the feature vector F_t^i and computes the kNN distance d_t^i with respect to the training instances in $\mathcal{F}^{M_2}$. The proposed algorithm then computes the instantaneous frame-level anomaly evidence δ_t:

$$\delta_t = (\max_i \{d_t^i\})^m - d_\alpha^m, \tag{2}$$

where m is the dimensionality of feature vector F_t^i. Finally, following a CUSUM-like procedure [2] we update the run-

ning decision statistic s_t as

$$s_t = \max\{s_{t-1} + \delta_t, 0\}, s_0 = 0. \quad (3)$$

For nominal data, δ_t typically gets negative values, hence the decision statistic s_t hovers around zero; whereas for anomalous data δ_t is expected to take positive values, and successive positive values of δ_t will make s_t grow. We decide that a video frame is anomalous if the decision statistic s_t exceeds the threshold h. After s_t exceeds h, we perform some fine tuning to better label video frames as nominal or anomalous. Specifically, we find the frame s_t started to grow, i.e., the last time $s_t = 0$ before detection, say τ_{start}. Then, we also determine the frame s_t stops increasing and keeps decreasing for n, e.g., 5, consecutive frames, say τ_{end}. Finally, we label the frames between τ_{start} and τ_{end} as anomalous, and continue testing for new anomalies with frame $\tau_{end} + 1$ by resetting $s_{\tau_{end}} = 0$.

Continual Learning: During testing, if the test statistic s_t at time t is zero, i.e, the feature vector F_t^i is considered nominal, then the feature vector is included in the second nominal training set $\mathcal{F}^{M_2}$. When the statistic s_t crosses the threshold h, an alarm is raised, signaling that the sequence of frames from τ_{start} to t have never occurred before in the training data. At this point, we propose a human-in-the-loop approach in which a human expert labels false alarms from time to time. If it is labeled as a false alarm, all vectors $\{F_\tau^i\}$ between τ_{start} and t are added to $\mathcal{F}^{M_2}$ so as to prevent similar future false alarms. Thanks to the kNN-based decision rule, such a sequential update enables the proposed framework to continually learn on recent data without the need for retraining from scratch, as opposed to the deep neural network-based decision rules.

3.6. Computational Complexity

In this section we analyze the computational complexity of the sequential anomaly detection module, as well as the average running time of the deep learning module.

Sequential Anomaly Detection: The training phase of the proposed anomaly detection algorithm requires the computation of kNN distance for each point in $\mathcal{F}^{M_1}$ with respect to each point in $\mathcal{F}^{M_2}$. Therefore, the time complexity of training phase is given by $\mathcal{O}(M_1 M_2 m)$. The space complexity of the training phase is $\mathcal{O}(M_2 m)$ since M_2 data instances need to be saved for the testing phase. In the testing phase, since we compute the kNN distances of a single point to all data points in $\mathcal{F}^{M_2}$, the time complexity is $\mathcal{O}(M_2 m)$. On the other hand, deep learning-based methods need to be retrained from scratch to avoid catastrophic forgetting, which would require them to store the old data as well as the new data. The space complexity of the deep learning-based methods would be $\mathcal{O}(abM_2)$ where $a \times b$ is the resolution of the video, which is typically much larger than m. Needles to say, the time complexity of retraining a deep learning-based detector is huge.

Deep Learning Module: The YOLO object detector requires about 12 milliseconds to process a single image. This translates to about 83.33 frames per second. Flownet 2 is able to process about 40 frames per second. Accounting for the sequential anomaly detection pipeline, the entire framework would approximately be able to process 32 frames per second. Hence, the proposed framework can process a surveillance video stream in real-time. We also report the running time for other methods such as 11 fps in [16] and 25 fps in [23]. The running time can be further improved by using a faster object detector such as YOLOv3-Tiny or a better GPU system. All tests are performed on NVIDIA GeForce RTX 2070 with 8 GB RAM and Intel i7-8700k CPU.

4. Experiments

4.1. Datasets

We first evaluate our proposed method on three publicly available benchmark video anomaly data sets, namely the CUHK avenue dataset [26], the UCSD pedestrian dataset [29], and the ShanghaiTech campus dataset [28]. Their training data consists of nominal events only. We present some examples of nominal and anomalous frames in Figure 2.

UCSD Ped2: The UCSD pedestrian data consists of 16 training and 12 test videos, each with a resolution of 240 x 360. All the anomalous events are caused due to vehicles such as bicycles, skateboarders and wheelchairs crossing pedestrian areas.

Avenue: The CUHK avenue dataset contains 16 training and 21 test videos with a frame resolution of 360 x 640. The anomalous behaviour is represented by people throwing objects, loitering and running.

ShanghaiTech: The ShanghaiTech Campus dataset is one of the largest and most challenging datasets available for anomaly detection in videos. It consists of 330 training and 107 test videos from 13 different scenes, which sets it apart from the other available datasets. The resolution for each video frame is 480 x 856.

4.2. Benchmark Algorithms

In the context of video anomaly detection, to the best of our knowledge, there is no benchmark algorithm designed for continual learning. Hence, in Table 1, we compare our proposed algorithm with the state-of-the-art deep learning-based methods, as well as methods based on hand-crafted features: MPPCA [18], MPPC + SFA [29], Del et al. [10], Conv-AE [13], ConvLSTM-AE [27], Growing Gas [41], Stacked RNN [28], Deep Generic [14], GANs [32], Liu et al. [23], Sultani et al. [40]. A popular metric used for com-

Figure 2. Examples of nominal and anomalous frames in the UCSD Ped2, CUHK Avenue and ShanghaiTech datasets. Anomalous events are shown with red box.

parison in the anomaly detection literature is the Area under the Curve (AuC) curve. Higher AuC values indicate better performance for an anomaly detection system. Following the existing works [7, 16, 23], we use the commonly used frame-level AuC metric for performance evaluation.

4.3. Impact of Sequential Anomaly Detection

To demonstrate the importance of sequential anomaly detection in videos, we implement a nonsequential version of our algorithm by applying a threshold to the instantaneous anomaly evidence δ_t, given in (2), which is similar to the approach employed by many recent works [23, 40, 16]. As Figure 3 shows, instantaneous anomaly evidence is more prone to false alarms than the sequential statistic of the proposed framework since it only considers the noisy evidence available at the current time to decide. Whereas, the proposed sequential statistic handles noisy evidence by integrating recent evidence over time.

4.4. Impact of Optical Flow

In Figure 4, we present the optical flow statistics for the first test video of the UCSD dataset. Here, the anomaly pertains to a person using a bike on a pedestrian path, which is previously unseen in the training data. It is clearly visible that there is a significant shift in the optical flow statistics, especially in skewness and kurtosis. This is due to the higher speed of a bike as compared to a person walking. Also, this shows the efficacy of optical flow in detecting motion-based anomalies.

4.5. Results

Benchmark Results: To show the general performance of the proposed algorithm, not necessarily with continual learning, we compare our results to a wide range of methods in Table 1 in terms of the commonly used frame-level AuC metric. Recently, [16] showed significant gains over the rest

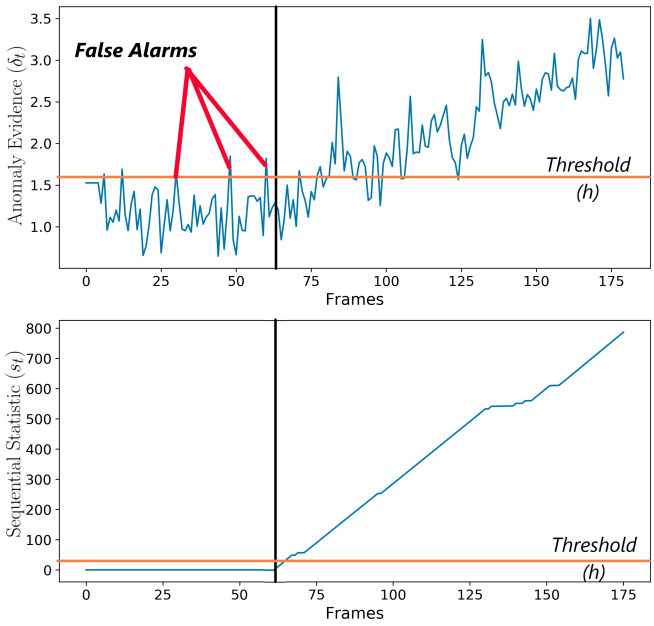

Figure 3. The advantage of sequential anomaly detection over single-shot detection in terms of controlling false alarms.

of the methods. However, their methodology of computing the AuC gives them an unfair advantage as they calculate the AuC for each video in a dataset, and then average them as the AuC of the dataset, as opposed to the other works which concatenate all the videos first and then determine the AuC as the datasets score.

As shown in Table 1, we are able to outperform the existing results in the CUHK Avenue and UCSD datasets, and achieve competitive performance in the ShanghaiTech dataset. We should note here that our reported result in the ShanghaiTech dataset is based on online decision making without seeing future video frames. A common technique

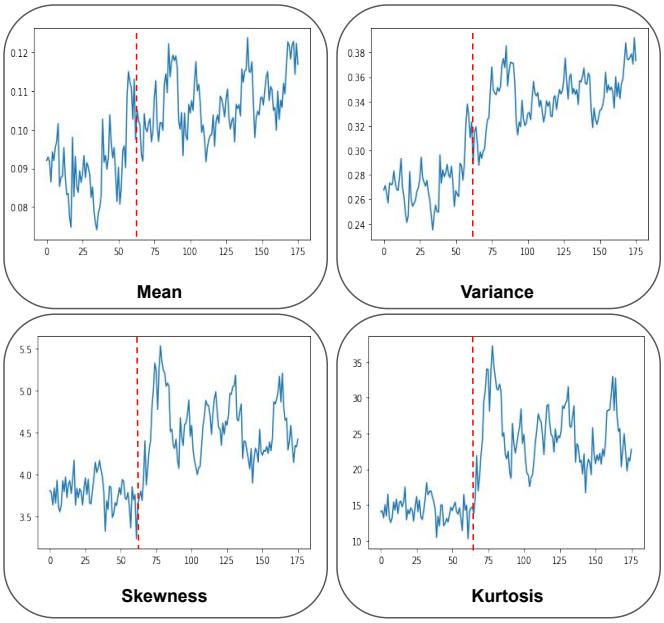

Figure 4. Optical flow statistics for the motion-based anomaly in the first test video of the UCSD Dataset.

Methodology	CUHK Avenue	UCSD Ped 2	ShanghaiTech
Conv-AE [13]	80.0	85.0	60.9
ConvLSTM-AE[27]	77.0	88.1	-
Stacked RNN[28]	81.7	92.2	68.0
GANs [32]	-	88.4	-
Liu et al. [23]	85.1	95.4	72.8
Sultani et al. [40]	-	-	71.5
Ours	86.4	97.8	71.62

Table 1. AuC result comparison on three datasets.

used by several recent works such as [23, 16] is to normalize the computed statistic for each test video independently using the future frames. However, this methodology cannot be implemented in an online (real-time) system as it requires prior knowledge about the minimum and maximum values the statistic might take.

Continual Learning Results: Due to the lack of existing benchmark datasets for continual learning in surveillance videos, we first slightly modify the original UCSD dataset, where a person riding a bike is considered as anomalous, and assume that it is considered as a nominal behavior. Our goal here is to compare the continual learning capability for video surveillance of the proposed and state-of-the-art algorithms and see how well they adapt to new patterns. Initially, the proposed algorithm raises an alarm when it detects a bike in the testing data. Using the human supervision approach proposed in Section 3, the relevant frames are labelled as nominal and added to the training set. In Figure 5, it is seen that the proposed algorithm clearly outperforms the state-of-the-art algorithms [16, 23] in terms of continual learning performance. More importantly, as shown in Table 2, it achieves this superior performance by quickly updating its training with the new samples in a few seconds while the state-of-the-art algorithms need to retrain on the entire dataset for several hours to prevent catastrophic forgetting. Furthermore, it is important to note that the proposed algorithm is able to achieve a relatively high AuC score using only a few samples, demonstrating its few-shot learning ability.

	Ours	Liu et al. [23]	Ionescu et al. [16]
Update Time	10 sec	4.8 hrs	2.5 hrs

Table 2. Time required to update the model for each batch of new samples.

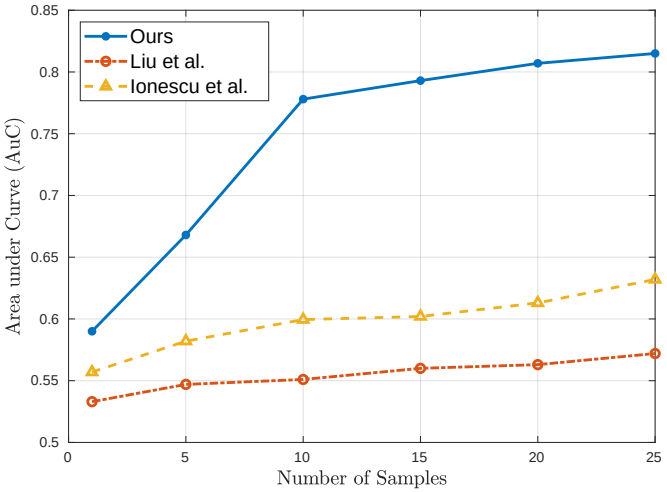

Figure 5. Comparison of the proposed and the state-of-the-art algorithms Liu et al. [23] and Ionescu et al. [16] in terms of continual learning capability. The proposed algorithm is able to quickly train with new samples and significantly outperform both of the methods.

Real-Time Surveillance Results: Even though existing datasets such as ShanghaiTech, CUHK Avenue, and UCSD provide a good baseline for comparing video surveillance frameworks, they lack some critical aspects. Firstly, they have an underlying assumption that all nominal events/behaviors are covered by the training data, which might not be the case in a realistic implementation. Secondly, there is an absence of temporal continuity in the test videos, i.e., most videos are only a few minutes long and there is no specific temporal relation between different test videos. Moreover, external factors such as brightness and weather conditions that affect the quality of the images are also absent in the available datasets. Hence, we also evaluate our proposed algorithm on a publicly available CCTV surveillance feed[1]. The entire feed is of 8 hours and 23 min-

[1]The entire surveillance feed is available here: https://www.youtube.com/watch?v=Xyj-7WrEhQwt=3460s

Figure 6. The visualization of different causes for false alarm in the surveillance feed dataset. In the first case, the person stands in the middle of the street, which causes an alarm as this behavior was previously unseen in the training data. Similarly, in the second case, a change in the weather causes the street sign to move. In the third case, the appearance of multiple cars at the same time causes a shift in the distribution of the optical flow. Finally, in the fourth case a bike is detected, which was not previously seen in the training data.

utes and continuously monitors a street. To make the problem more challenging we initially train only on 10 minutes of data and then continually update our model as more instances become available.

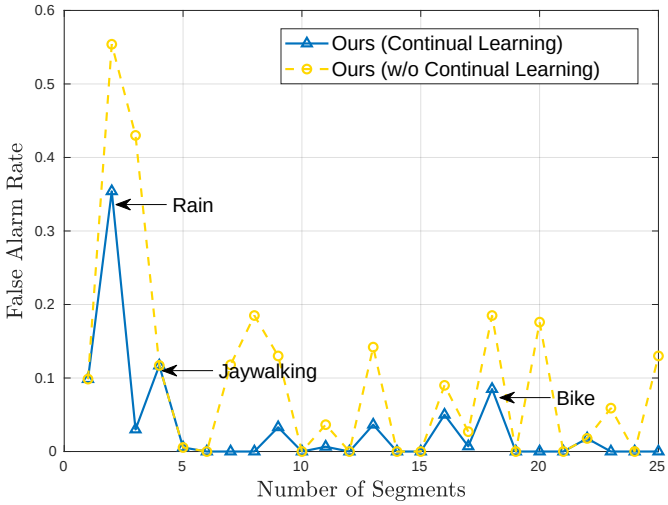

Figure 7. Continual learning ability of the proposed algorithm. Assuming an arbitrary constant threshold, we observe that the algorithm is able to quickly learn new nominal behaviors, and thus reduce the false alarm rate as compared to the same algorithm which does not continually update the learned model.

In Figure 7, we demonstrate the continual learning performance of the proposed algorithm through reduced number of false alarms after receiving some new nominal labels. It should be noted that our goal here is to emphasize the continual learning ability of our algorithm, rather than showing the general detection performance. Each segment here corresponds to 20,000 frames. After each segment, a human roughly labels the false positive events. In this case, to reduce the computational complexity and examine the few-shot learning ability of the proposed algorithm, we only consider 20% of all false positive events for updating the model. Although the number of frames might seem a lot, it roughly translates to 10 seconds of streaming video data, so it can still be considered as few-shot learning in video analysis. We observe that even with relatively small updates, the false alarm rate is significantly lower as compared to the same algorithm where we do not update the model continuously. This proves that the proposed algorithm is able to learn meaningful information from recent data using only few samples, and is able to incrementally update the model without accessing the previous training data.

5. Conclusion and Future Work

For video anomaly detection, we presented an continual learning algorithm which consists of a transfer learning-based feature extraction module and a statistical decision making module. The first module efficiently minimizes the training complexity and extracts motion, location, and appearance features. The second module is a sequential anomaly detector which is able to incrementally update the learned model within seconds using newly available nominal labels. Through experiments on publicly available data, we showed that the proposed detector significantly outperforms the state-of-the-art algorithms in terms of any-shot learning of new nominal patterns. The continual learning capacity of the proposed algorithm is illustrated on a real-time surveillance stream, as well as a popular benchmark dataset.

The ability to continually learn and adapt to new scenarios would significantly improve the current video surveillance capabilities. In future, we aim to evolve our framework to work well in more challenging scenarios such as dynamic weather conditions, rotating security cameras and complex temporal relationships. Furthermore, we plan to extend the proposed continual learning framework to new anomalous labels, and other video processing tasks such as online object and action recognition.

References

[1] Nadeem Anjum and Andrea Cavallaro. Multifeature object trajectory clustering for video analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1555–1564, 2008. 2

[2] Michèle Basseville and Igor V Nikiforov. *Detection of abrupt changes: theory and application*, volume 104. prentice Hall Englewood Cliffs, 1993. 4

[3] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1932–1939. IEEE, 2009. 2

[4] George H Chen, Devavrat Shah, et al. Explaining the success of nearest neighbor methods in prediction. *Foundations and Trends® in Machine Learning*, 10(5-6):337–588, 2018. 1, 4

[5] Kai-Wen Cheng, Yie-Tarng Chen, and Wen-Hsien Fang. Video anomaly detection and localization using hierarchical feature representation and gaussian process regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2909–2917, 2015. 2

[6] Rensso Victor Hugo Mora Colque, Carlos Caetano, Matheus Toledo Lustosa de Andrade, and William Robson Schwartz. Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(3):673–682, 2016. 2

[7] Yang Cong, Junsong Yuan, and Ji Liu. Sparse reconstruction cost for abnormal event detection. In *CVPR 2011*, pages 3449–3456. IEEE, 2011. 6

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016. 1

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005. 2

[10] Allison Del Giorno, J Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In *European Conference on Computer Vision*, pages 334–349. Springer, 2016. 5

[11] Zhouyu Fu, Weiming Hu, and Tieniu Tan. Similarity based vehicle trajectory clustering and anomaly detection. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–602. IEEE, 2005. 2

[12] Xiaoyi Gu, Leman Akoglu, and Alessandro Rinaldo. Statistical analysis of nearest neighbor methods for anomaly detection. In *Advances in Neural Information Processing Systems*, pages 10921–10931, 2019. 1, 4

[13] Mahmudul Hasan, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis. Learning temporal regularity in video sequences. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 733–742, 2016. 2, 5, 7

[14] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3619–3627, 2017. 2, 5

[15] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 4

[16] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7842–7851, 2019. 1, 2, 5, 6, 7

[17] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in neural information processing systems*, pages 5541–5552, 2018. 1, 2

[18] Jaechul Kim and Kristen Grauman. Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2928. IEEE, 2009. 5

[19] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 1, 3

[20] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015. 2

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[22] L. Lin and N. Purnell. A world with a billion cameras watching you is just around the corner. *The Wall Street Journal, https://www.wsj.com/articles/a-billion-surveillance-cameras-forecast-to-be-watching-within-two-years-11575565402*, 2019. 1

[23] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6536–6545, 2018. 1, 2, 5, 6, 7

[24] Vincenzo Lomonaco and Davide Maltoni. Comparing incremental learning strategies for convolutional neural networks. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 175–184. Springer, 2016. 2

[25] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. *arXiv preprint arXiv:1705.03550*, 2017. 2

[26] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 5

[27] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 439–444. IEEE, 2017. 2, 5, 7

[28] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017. 2, 5, 7

[29] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. IEEE, 2010. 5

[30] Huizi Mao, Xiaodong Yang, and William J Dally. A delay metric for video object detection: What average precision fails to tell. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 573–582, 2019. 1, 3

[31] Romero Morais, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11996–12004, 2019. 2

[32] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-play cnn for crowd motion analysis: An application in abnormal event detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1689–1698. IEEE, 2018. 5, 7

[33] Mahdyar Ravanbakhsh, Enver Sangineto, Moin Nabi, and Nicu Sebe. Training adversarial discriminators for cross-channel abnormal event detection in crowds. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1896–1904. IEEE, 2019. 2

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 3

[35] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016. 1

[36] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3379–3388, 2018. 2

[37] Venkatesh Saligrama and Zhu Chen. Video anomaly detection based on local statistical aggregates. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2112–2119. IEEE, 2012. 2

[38] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017. 1

[39] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pages 4077–4087, 2017. 2

[40] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018. 1, 2, 5, 6, 7

[41] Qianru Sun, Hong Liu, and Tatsuya Harada. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition*, 64:187–201, 2017. 5

[42] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 2

[43] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *Advances in neural information processing systems*, pages 3630–3638, 2016. 2

[44] Dan Xu, Elisa Ricci, Yan Yan, Jingkuan Song, and Nicu Sebe. Learning deep representations of appearance and motion for anomalous event detection. *arXiv preprint arXiv:1510.01553*, 2015. 2