Any-Shot Sequential 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. Even though the performance of state-of-the-art methods on publicly available data sets has been competitive, they demand a massive amount of training data. Also, they lack a concrete approach for continuously updating the trained model once new data is available. 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 any-shot learning, which in turn significantly reduces the training complexity and provides a mechanism which can detect anomalies using only a few labeled nominal examples. Our proposed algorithm leverages the feature extraction power of neural network-based models for transfer learning, and the any-shot learning capability of statistical detection methods.

1. Introduction

The rapid advancements in the technology of closed-circuit television (CCTV) cameras and their underlying infrastructure has led to a sheer number of surveillance cameras being implemented globally, estimated to go beyond 1 billion by the end of the year 2021 [17]. Considering the massive amounts of videos generated in real-time, manual video analysis by human operator becomes inefficient, expensive, and nearly impossible, which in turn makes a great demand for automated and intelligent methods for an efficient video surveillance system. An important task in video surveillance is anomaly detection, which refers to the identification of events that do not conform to the expected behavior [2].

A vast majority of the recent video anomaly detection methods directly depend on deep neural network architectures [32]. It is well known that these deep neural network models are data hungry. As a result, they require many labeled nominal frames and long hours of training to produce acceptable results on a new data set. Moreover, most of them are not suitable for online detection of anomalies as they need the knowledge of future video frames for appropriate normalization of detection score.

Motivated by the aforementioned domain challenges and research gaps, we propose a hybrid use of neural networks and statistical $k$ nearest neighbor ($k$NN) decision approach for finding video anomalies with limited training in an online fashion. In summary, our contributions in this paper are as follows:

- We significantly reduce the training complexity by leveraging transfer learning while simultaneously outperforming the current state-of-the-art algorithms.
- We propose a novel framework for statistical any-shot sequential anomaly detection which is capable of learning continuously and from few samples.
- Extensive evaluation on publicly available data sets show that our proposed framework can transition effectively between few-shot and many-shot learning.

2. Related Works

Anomaly detection methods for video surveillance can be broadly classified into two categories: traditional and deep learning-based methods. Traditional methods [6, 21, 29, 37] extract hand-crafted motion and appearance features such as histogram of optical flow [3, 4] and histogram of oriented gradients [6] to detect spatiotemporal anomalies [29]. The recent literature is dominated by the neural network-based methods [8, 9, 18, 20, 26, 28, 35] due to their superior performance [35]. For instance, in [19], Convolutional Neural Networks (CNN), and Convolutional Long Short Term Memory (CLSTM) are used to learn appearance and motion features. More recently, Generative Adversarial Networks (GANs) have been used to generate internal scene representations based on a given frame and its optical flow to detect deviation of the GAN output from the nominal data [18, 25]. However, there is also a significant debate on the shortcomings of neural network-based
methods in terms of interpretability, analyzability, and reliability of their decisions [12]. For example, [23, 30] propose using a nearest neighbor-based approach together with deep neural network structures to achieve robustness, interpretability for the decisions made by the model, and defense against adversarial attack. Also, deep neural networks for visual recognition typically require a large amount of labelled examples for training [16], which might not be available for all possible behaviors/patterns. Hence, recently researchers have begun to address the challenge of few-shot learning [14, 31, 33, 34]. A line of few-shot learning methods is based on the idea of transfer learning, i.e., using a pre-trained model learned from one domain for another domain [15, 22, 36].

3. Proposed Method

An anomaly is construed as an unusual event which does not conform to the learned nominal patterns. However, in general, for practical implementations, it is unrealistic to assume the availability of sufficient training data for all possible nominal patterns/events. Thus, a practical framework should be able to perform any-shot learning of nominal events. This presents a novel challenge to current approaches mentioned in Section 2 as their decision mechanism is extensively dependent on Deep Neural Networks (DNNs). DNNs typically require a large amount of training data with sufficient number of samples for each type of nominal event or exhibit the risk of catastrophic forgetting [13]. Also, in general, the type of anomaly that the detector might encounter 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 $x^t_i$ for each object $i$ in frame $X^t$ at time $t$, where $x^t_i$ is given by $[w_1 x_{\text{motion}}, w_2 x_{\text{location}}, w_3 x_{\text{appearance}}]$. The weights $w_1, w_2, w_3$ are used to adjust the relative importance of each feature category and are set as $[0.9, 0.5, 1]$ respectively.

3.1. 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 continuously learn. Hence, we propose to leverage transfer learning to extract meaningful features from video.

Object Detection: To obtain location and appearance features, we propose to detect objects using a pre-trained real-time object detection system such as You Only Look Once (YOLO) [27]. YOLO offers a higher frames-per-second (fps) processing while providing better accuracy as compared to the other state-of-the-art models such as SSD and ResNet. For online anomaly detection, speed is a critical factor, and hence we currently prefer YOLOv3 trained on the MS COCO dataset in our implementations. For each detected object in image $X_t$, we get a bounding box (location) along with the class probabilities (appearance). 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 3.2.

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 [10]. 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.

Combining the motion, location, and appearance features, for each object detected in a frame, we construct a feature vector as shown in Fig. 1, where Mean, Variance, Skewness and Kurtosis are extracted from the optical flow; $C_x, C_y, Area$ denote the coordinates of the center of the bounding box and the area of the bounding box (Section 3.1); and $p(C_1), \ldots, p(C_n)$ are the class probabilities for the detected object (Section 3.1). Hence, at any given time $t$, with $n$ denoting the number of possible classes, the dimensionality of the feature vector is given by $D = n + 7$.

3.2. Any-Shot Sequential Anomaly Detection

Anomaly detection in streaming video fits well to the sequential change detection framework [1] as we can safely assume that any anomalous event would persist for an unknown but significant period of time. The eventual goal is to detect anomalies with minimal detection delays while satisfying a desired false alarm rate. Traditional parametric change detection algorithms which require probabilistic models cannot be implemented directly here as no prior knowledge about the anomalous events is available. Moreover, it is unrealistic to assume that the available training data includes sufficient number of frames for every possible nominal event. For example, while monitoring a street, the number of frames available for a car would be much more than a truck.

Training: In the N-shot video setting, given a set of $N$ nominal frames, we leverage our transfer learning module to extract the training set $S_M = \{x_1, \ldots, x_M\}$, where $M$ is the number of detected objects, and $x_i \in \mathbb{R}^D$ is a $D$-
Figure 1. Proposed few-shot 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 automatically sets its decision threshold to satisfy a false alarm constraint and makes online decisions.

dimensional feature vector. Assuming that the training data does not include any anomalies, $\{x_1, \ldots, x_M\}$ correspond to $M$ points in the nominal data space, distributed according to an unknown complex probability distribution. To determine the nominal data patterns in a nonparametric way, we use $k$-nearest-neighbor ($k$NN) Euclidean distance to capture the interactions between the nominal data points due to its essential traits such as analyzability, interpretability, and computational efficiency, which deep learning-based models sorely lack. Given the informativeness of extracted motion, location, and appearance features, anomalous instances are expected to lie further away from the nominal training (support) set, which will lead to statistically higher $k$NN distances for the anomalous instances in the test (query) set with respect to the nominal data points. The training procedure of our detector is given as follows:

1. Normalize the training set $S_M$ using the min-max approach and partition it into two sets $S_{M_1}$ and $S_{M_2}$ such that $M = M_1 + M_2$, where $M_1$ and $M_2$ are the sizes of the partitions respectively. One way is to select the first $M_1$ instances for $S_{M_1}$ and the remaining instances for $S_{M_2}$ such that $M_1 < M_2$.

2. Then for each feature vector $x_i$ in $S_{M_1}$, we compute the $k$NN distance $d_i$ with respect to the points in $S_{M_2}$.

3. For a significance level $\alpha$, e.g., 0.05, the $(1 - \alpha)$th percentile $d_\alpha$ of $k$NN 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 sequential anomaly detection algorithm constructs the feature vector $x_i^t$ and computes the $k$NN distance $d_i^t$ with respect to the training instances in $S_{M_2}$. Then, the instantaneous frame-level anomaly evidence $\delta^t$ is computed as

$$\delta^t = (\max_i d_i^t)^D - (d_\alpha)^D.$$  \hspace{1cm} (1)

Finally, following a CUSUM-like procedure [1] we update the running decision statistic $s_t$ as

$$s_t = \max\{s_{t-1} + \delta^t, 0\}, s^0 = 0.$$  \hspace{1cm} (2)

We decide that there exists an anomaly in video if the decision statistic $s_t$ exceeds the threshold $h$. After the anomaly decision, to determine the anomalous frames, we find the frame $s^t$ started to grow, say $\tau_{\text{start}}$, and also determine the frame $s^t$ stops increasing and keeps decreasing for a certain number, e.g., 5, of consecutive frames, say $\tau_{\text{end}}$. Finally, we label the frames between $\tau_{\text{start}}$ and $\tau_{\text{end}}$ as anomalous, and continue testing for new anomalies with frame $\tau_{\text{end}} + 1$ by resetting $s_{\tau_{\text{end}}} = 0$.

Existing works consider the decision threshold $h$ as a design parameter, however for a practical anomaly detection algorithm, a clear procedure for selecting it is necessary. In [7], we provide an asymptotic ($M_2 \rightarrow \infty$) upper bound on the false alarm rate:

$$\text{FAR} \leq e^{-\omega_0 h},$$  \hspace{1cm} (3)
where \( \omega_0 > 0 \) is given by

\[
\omega_0 = v_m - \theta - \frac{1}{\phi} W \left( -\phi \theta e^{-\phi \theta} \right), \tag{4}
\]

\[
\theta = \frac{v_m}{e^{v_m d_m}}.
\]

In (4), \( W(\cdot) \) is the Lambert-W function, \( v_m = \frac{\pi^{m/2}}{\Gamma((m/2)+1)} \) is the constant for the \( m \)-dimensional Lebesgue measure (i.e., \( v_m d_m^m \) is the \( m \)-dimensional volume of the hyperball with radius \( d_m \)), and \( \phi \) is the upper bound for \( \delta^t \). Although the expression for \( \omega_0 \) looks complicated, all the terms in (4) can be easily computed. Particularly, \( v_m \) is directly given by the dimensionality \( m \), \( d_m \) comes from the training phase, \( \phi \) is also found in training, and finally there is a built-in Lambert-W function in popular programming languages such as Python and Matlab. Hence, given the training data, \( \omega_0 \) can be easily computed, and the threshold \( h \) can be chosen to asymptotically achieve the desired false alarm period as follows

\[
h = -\frac{\log(FAR)}{\omega_0}. \tag{5}
\]

4. Experiments

Most of the recent works evaluate their performance on three publicly available benchmark data sets, namely the UCSD pedestrian data set, the ShanghaiTech campus data set and the CUHK avenue data set. Even though each data set has its own set of challenges, all the data sets have a common nominal definition. This makes them susceptible to trivial algorithmic designs which can achieve competitive results as there is a very obvious shift between the nominal and anomalous distributions. Hence, to make the problem more challenging and test the any-shot learning capabilities of different state-of-the-art algorithms, we also test on a modified version of the UCSD data set. For performance evaluation, following the existing works [5, 11, 18], we use the frame-level Area under the Curve (AuC) metric.

**Any-shot learning:** As compared to the original UCSD data set, where a person riding a bike is considered as anomalous, in this case we assume that it is a nominal behavior with very few training samples. However, the remaining anomalous events occurring in the dataset such as a person riding a skateboard or a cart passing through are still considered as anomalous. Our goal here is to compare the any-shot learning capability of the proposed and state-of-the-art algorithms and see how well they adapt to new patterns. In this case, in addition to the available training data, we also train on a few samples of a person riding a bike. In Fig. 2, it is seen that the proposed algorithm clearly outperforms the state-of-the-art algorithms [11, 18] in terms of any-shot learning performance. It is important to note that for video applications, 10 shots correspond to less than a second in real time.

![Figure 2. Comparison of the proposed and state-of-the-art algorithms Liu et al. [18] and Ionescu et al. [11] in terms of any-shot learning. The proposed algorithm is able to transition well between few-shot and many-shot learning.](image)

<table>
<thead>
<tr>
<th>Methodology</th>
<th>CUHK Avenue</th>
<th>UCSD Ped 2</th>
<th>ShanghaiTech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv-AE [6]</td>
<td>80.0</td>
<td>85.0</td>
<td>60.9</td>
</tr>
<tr>
<td>ConvLSTM-AE [19]</td>
<td>77.0</td>
<td>88.1</td>
<td>-</td>
</tr>
<tr>
<td>Stacked RNN [20]</td>
<td>81.7</td>
<td>92.2</td>
<td>68.0</td>
</tr>
<tr>
<td>GANs [25]</td>
<td>-</td>
<td>88.4</td>
<td>-</td>
</tr>
<tr>
<td>Liu et al. [18]</td>
<td>85.1</td>
<td>95.4</td>
<td>72.8</td>
</tr>
<tr>
<td>Ours</td>
<td>86.4</td>
<td>97.8</td>
<td>71.62</td>
</tr>
</tbody>
</table>

Table 1. AuC result comparison on three datasets.

**Benchmark Datasets:** To show the competitive performance of the proposed algorithm with large training data, we compare our results on the entire data sets to a wide range of methods in Table 1. We should note here that our reported result in all the data sets is based on online decision making without seeing future video frames. A common technique used by several recent works [11, 18] is to normalize the computed statistic for each test video independently, which is not suitable for online detection.

5. Conclusion

For video anomaly detection, we presented an online anomaly detection 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 enables a clear procedure for selecting decision threshold through asymptotic performance analysis. 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.
References