Anomaly Detection in Surveillance Videos

Sukalyan Bhakat
Indian Institute of Technology Bombay
Mumbai, Maharashtra
sbhakat16@cse.iitb.ac.in

Ganesh Ramakrishnan
Indian Institute of Technology Bombay
Mumbai, Maharashtra
ganesh@cse.iitb.ac.in

ACM Reference format:
Sukalyan Bhakat and Ganesh Ramakrishnan. 2019. Anomaly Detection in Surveillance Videos. In *Proceedings of 6th ACM IKDD CoDS and 24th COMAD, Kolkata, India, January 3–5, 2019 (CoDS-COMAD '19),* 4 pages.
https://doi.org/10.1145/3297001.3297034

1 ABSTRACT

Automated anomaly detection is a useful task that can aid investigations and detect crimes. To this end, we present a model that can be used as a tool for anomaly detection in surveillance videos. Following an unsupervised approach, we use an autoencoder model trained to minimize the reconstruction error between the input and the generated output. We also augment the training of the auto-encoder with supervision in the form of user ratings per frame; higher user ratings reflect normal behaviour that the model is expected to faithfully reconstruct. On the other hand, lower rated frames are suspected to be anomalous. We analyze the output of the autoencoder on a standard dataset as well as two of our datasets that we have made public. We study the behavior of reconstruction error with and without supervision as well as the temporal coherence of the reconstruction error. Additionally, we use Grad-CAM to highlight potentially anomalous regions in the input. Finally, we discuss the problem of constructing summaries based on anomalous segments using heuristic approaches as well as a graph-theoretic formulation of determining a ranked list of maximum weighted cliques. We also make available in a single tool, our auto-encoder model as well as the anomaly summarizer.

2 INTRODUCTION

Surveillance is an integral part of any nation. CCTV cameras are ubiquitous and are used at various places. A system having the capability to detect and report suspicious activities is desirable and important. But, an event that is abnormal in one context may not be suspicious in another setting. Also, the anomalies are characterized by their properties in both the spatial as well as the temporal domain.

Taking the above challenges into consideration, we seek an unsupervised approach and propose an autoencoder model [2] for the task of anomaly detection. To the best of our knowledge, this is the first work to combine anomaly detection that allows unsupervised and semi-supervised training and includes Grad-CAM for analyzing the workings of the model. Moreover, videos depicting the anomalous segments are generated as output. We present the average *AUC of ROC curves* for all the three datasets; Avenue, office and police datasets. In order to supplement our claims, we also attach screenshots of the anomalies detected by our method and share a link, which contains illustrative results.

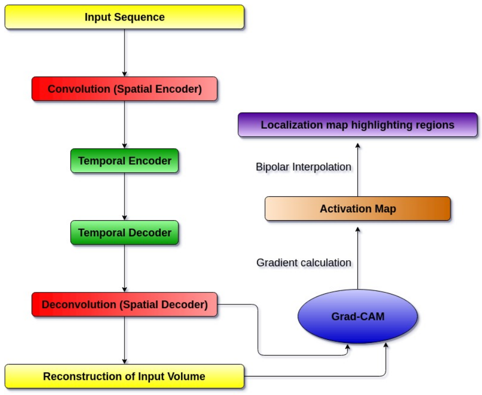
Input Sequence
Convolution (Spatial Encoder)
Temporal Encoder
Temporal Decoder
Deconvolution (Spatial Decoder)
Reconstruction of Input Volume
Grad-CAM
Gradient calculation
Activation Map
Bipolar Interpolation
Localization map highlighting regions

Figure 1: Model Architecture

3 RELATED WORK

Anomaly Detection: Anomaly detection is an exciting and challenging problem in the domain of computer vision. Various approaches such as clustering techniques, aggressive action detection techniques and tracking methods [7] have been proposed. For crowded scenes, global motion patterns are learnt using techniques like histogram-based methods [8] and topic modelling [6]. Dynamic Bayesian Networks and HMMs encode dependencies between variables and build a relationship between them.

In recent times, autoencoders [3, 4, 9, 10] have come into use that learn the patterns of normal behaviour and detect abnormalities by calculating a deviation between the input and the output. Generally, they do not require annotated data and can be trained in an unsupervised manner. Our work is motivated by the work done in [2], which calculates the reconstruction error at test time in order to tag frames as normal or anomalous.

Grad-CAM analysis: Gradient Class Activation Mapping is a technique to highlight regions of the input that led to the decision made by the model. [5] uses gradients of a target class flowing into the output layer to form a localization map for three tasks namely image classification, image captioning and visual question answering.

4 DATASETS

Three datasets, namely *Avenue, Surveillance office* and *Police* have been used. Experiments have been conducted on all the three datasets.

CoDS-COMAD '19, January 3–5, 2019, Kolkata, India

ACM ISBN 978-1-4503-6207-8/19/01.
https://doi.org/10.1145/3297001.3297034

We selected the standard 'Avenue' dataset for comparing our results with the existing work. The other two datasets are our own datasets, which have been created to challenge our models with more real and diverse anomalous situations. They have been named *Surveillance Office* and *Police*, respectively.

4.1 Avenue dataset

It consists of 16 training (15328 frames) and 21 testing (15324 frames) videos. The videos have been captured at the CUHK (Chinese University of Hong Kong) campus [1]. Normal scenes consist of people moving in the background. Anomalies in the testing videos include i) strange action, ii) wrong direction and iii) abnormal objects.

4.2 Surveillance office dataset

It is one of our original datasets and consists of 16 training (637712 frames) and 3 testing (100238 frames) videos. The training videos contain normal activities such as i) meetings and ii) people moving around inside the office and span around 30 minutes in duration. The testing videos contains anomalies like i) fighting, ii) agitated movements, iii) carrying strange objects and iv) doing suspicious actions like covering the camera.

4.3 Police dataset

This dataset contains 357 training (180051 frames) and 5 testing (14929 frames) videos. The videos depict a scene of a booth where people withdraw money. The training videos showcase either i) an empty booth or ii) people withdrawing money. The testing videos contain anomalous events such as i) a person wearing a helmet inside the booth, ii) a dog inside the booth and iii) a person stealing money. Since, the surveillance and police datasets are original, we create the ground-truth and rating files using a software called *oTranscribe*.

5 PROPOSED APPROACH

We utilize the idea of spatio-temporal autoencoders and exploit their ability to learn patterns of normal events without any supervision. We explain the architecture of our model followed by the approach and conclude the section with some results. As an enhancement, we introduce user ratings and include a parameter for the same in the objective function that is minimized during training. We discuss the updated model and present revised results on all the three datasets. Finally, we use the analysis method Grad-CAM for gaining useful insights into the inner workings of our model. [5] used Grad-CAM technique with respect to a target class. But we utilize this method to analyze the output irrespective of any class. We explain the process mathematically and follow it up with results in form of images that depict the highlighted regions.

5.1 Autoencoder model

This section describes the autoencoder model, which is fully unsupervised and utilizes both the spatial as well as the temporal characteristics of a video to detect anomalies during test time. During training, the model tries to learn the characteristics of normal events by minimizing the reconstruction error associated with each frame.

Volumes, comprising of t frames, are input to the model. These volumes impart a temporal aspect as the model sees a sequence of frames at a time. The value of t is empirical. Larger the value of t, greater will be the temporal context that can be exploited by the model.

5.1.1 Architecture. Autoencoder is a neural network that consists of 2 stages namely the encoding stage and the decoding stage, as shown in figure 1. The model accepts a volume as an input and attempts to recreate it on the output side. The objective is to minimize the reconstruction error. Annotated data is not required during training time.

A volume is fed as input. Each volume consists of t frames, each with dimensions 224 x 224. Two convolution layers extract the spatial characteristics followed by two convolution LSTMs that capture the temporal features. The encoder module is mirrored on the decoder side in form of a convolution LSTM that is followed by two deconvolution layers to reconstruct the output. The output layer is the reconstruction of the input. Hence, dimension of this layer is same as the input layer.

5.1.2 Training Phase. The training phase consists of a data preprocessing step that operates on the video frames to prepare the input in the desired format. Mean value of all the frames is calculated and subtracted from individual frames for normalization. Volumes are also created and the model is trained for a given number of epochs. The training videos should consist of normal events only. Mean-squared error between the reconstructed frame and the original frame is minimized during the training process using a gradient descent optimizer.

5.1.3 Testing Phase. Volumes are fed as input and the output is a reconstructed version of the entire volume. We desire a score for each frame but the output of the model is the reconstruction error of a volume. So, we consider each reconstructed volume as the representative of a frame. We maintain batches for faster execution. The reconstruction error is written to a file, which is later used for analysis and creation of summaries.

5.2 Semi-supervised autoencoder model

Anomaly detection is a context-dependent task and receiving external aid during training is always desirable. Hence, we permit users to rate segments of the videos. Segments rated 1 reflect normal behaviour, which the model is expected to reconstruct faithfully and the frames rated -1 are anomalous. We manually go through the videos of our dataset and rate segments using an open source tool called $oTranscribe$. The rating files are also available in the dataset repository, which we have made public.

5.2.1 Drawbacks of the autoencoder model. The model described in the section 5.1 is robust but there are two major drawbacks.

Training data: It is imperative that training dataset contains videos consisting of normal events only. Segregation of videos into training and testing set requires lot of human effort and time. Thus, it is desirable to have a model that does not pose restrictions on the content of videos available for training.

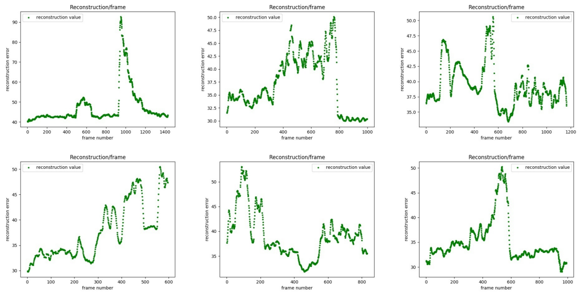

Figure 2: Plot of reconstruction error vs frame number. The peaks denote a region of interest as the model predicts an output having a large reconstruction error.

Blind training: A user does not have the means to convey that a seemingly normal event could potentially lead to something suspicious.*Eg:* a person carrying a knife looks like a person carrying a cuboid box, but both the events could lead to very different situations.

5.2.2 Updated model. In order to address some of the issues mentioned in subsection 5.2.1, we introduce user ratings and update the model to accommodate the changes. Frames corresponding to the normal and abnormal events are rated 1 and -1, respectively. The objective of the model described in section 5.1 was to minimize the reconstruction error value during training. This was correct because the training data consisted of videos containing normal events only.

$$Objective = \forall_{i \epsilon frames} \min(L_i) \tag{1}$$

Now, we add a parameter corresponding to user ratings into the objective function. It is a multiplicative objective function where we want to minimize the reconstruction error multiplied by the user rating.

$$Objective = \forall_{i \epsilon frames} \min(r_i * L_i) \tag{2}$$

When a frame is normal, r_i is 1 and the model tries to minimize L_i. On the other hand, when a frame is anomalous, r_i is -1 and the model adjusts its weights in such a way that the output is not reconstructed properly.

Now, we can have a mix of normal and abnormal videos in the training set. L_i will be minimized for normal events and maximized for abnormal events.

6 EVALUATION

We present 3 different settings for evaluation. Classification, plot of reconstruction error versus frame number (as shown in fig 2) and anomaly summarizers.

6.1 Thresholding

A threshold value is calculated with respect to the reconstruction error. Frames with a reconstruction error greater than the threshold are classified as anomalous and vice-versa.

$$Threshold = MEDIAN - (k * SD) \tag{3}$$

SD is the Standard Deviation and k is a constant that can be set empirically. We assume that at least 50 percent frames denote normal behaviour. With a lower threshold value, we ensure that all the anomalies are caught, but that might increase the number of *False Positives*. For comparison, we have reported the average AUC of the ROC curves for all the three datasets (see table 1).

AUC: without user ratings		
Dataset	Threshold	Sorted
Avenue	0.719	0.722
Office	0.641	0.624
Police	0.758	0.738

Table 1: Average AUC of ROC for all datasets

6.2 Sorting and selecting top p

We sort the frames of a test video in the decreasing order of their reconstruction error. The most anomalous frame will be present in the beginning of the sorted list. The top p percent frames are classified as anomalous. The value of p determines the size of the abnormal frames list. Also, we report the average AUC of the ROC curves for all the datasets (see table 2).

AUC: with user ratings		
Dataset	Threshold	Sorted
Avenue	0.725	0.759
Office	0.680	0.672
Police	0.762	0.791

Table 2: Average AUC of ROC for all datasets

We find that the average AUC value for the rated autoencoder is slightly greater than the average AUC value for the fully unsupervised one. This is expected as the inclusion of ratings aids the training process and enhances the performance of the model.

6.3 Anomaly summarizers

We create summary videos that depict the anomalous regions of a test video. For the creation of summaries, three different methods are proposed; i) thresholding, ii) sorting-selection and iii) a graph-theoretic formulation.

i) We set a threshold value for identifying anomalous frames and concatenate them in order to form the output video.

ii) We sort the frames according to the decreasing value of the reconstruction error. The top p percent frames are classified as anomalous and a fixed neighbourhood is concatenated to form the output video. The idea of picking a fixed neighbourhood arose from the insights we gathered after studying the reconstruction error vs frame number plot as shown in figure 2. As the anomaly score peaks multiple times over the duration of a video, we realize that the anomaly function displays a smooth behavior.

iii) Finally, we propose a graph-theoretic approach for anomaly summarization that would enable variable-sized windows. Attempt is to convert the premise of anomalies and the neighbourhood of frames into a graph problem. The construction of the graph is as follows:

- **Vertices -** anomalous frames, which are decided by setting a threshold based on the reconstruction error. The threshold can be empirical or calculated based on a formula as mentioned in 6.1. Vertex weight is the reconstruction error of the frame that is represented by the corresponding vertex.

- **Edges -** frames at a distance lesser than a threshold distance have edges between them. The threshold can be empirical. Edge weight is the magnitude of the distance between the two frames (vertices) that are adjacent.

Once we obtain a graph, we find the maximum weighted clique. This operation ensures that we select a subset of the frames that have the maximum reconstruction error, yet are nearby each other. A maximum weighted clique resembles a segment of the output video. Once, we obtain a clique, we remove that portion of the graph and look for the next maximum weighted clique. In this way, diversity of the anomalies is maintained.

We know that there is no efficient algorithm for finding the maximum weighted clique. But, if we select all the vertices and set a threshold for deciding the edges, the problem of finding the maximum weighted clique boils down to sliding a window across the duration of the video and selecting that segment, which has the maximum reconstruction error. This simplified problem can be solved in time complexity of $O(n)$, which is linear in the input size.

Anomaly summarization illustrates example summaries that have been generated by the methods discussed. Please refer to the supplementary materials for illustrations. Also, a working code is available online.

7 ANALYSIS

Deep learning models improve performance at the cost of transparency. In our case, we are unable to explain why a particular frame is assigned a higher reconstruction error value. In order to address this problem, we use a technique known as Grad-CAM or Gradient-based Class Activation Mapping. Grad-CAM has been used after the testing phase. It uses the existing gradients for calculations and is a one step process. As output, we get a grey-scale version of the input image with specific regions highlighted as shown in figure 3.

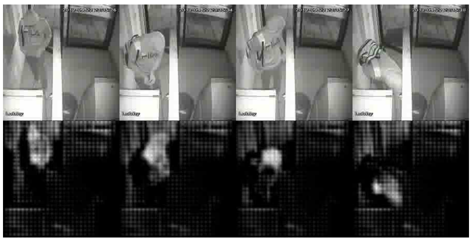

Figure 3: A person stealing money from an ATM booth. Grad-CAM highlights the head region

7.1 Grad-CAM

This section explains the theory behind Grad-CAM and how we have used it. For illustration, we consider the output layer and the penultimate layer, which is a deconvolution layer (as shown in fig 1).

7.1.1 Theory. Grad-CAM uses the gradients flowing between layers to create a localization map.

$$\alpha_k = \frac{1}{Z} \sum_{i,j,k,l} \frac{\delta y}{\delta A^k} \tag{4}$$

α_k is a scalar that denotes the importance of the k^{th} filter map. It is calculated by evaluating the gradient of the output layer with respect to the penultimate layer and normalizing along all the dimensions. i, j, k, l denote the dimensions.

$$L_{Grad-CAM} = ReLU\left(\sum_k \alpha_k A^k\right) \tag{5}$$

The importance score is multiplied with the convolution filter maps of the penultimate layer and aggregated to form a map with dimensions similar to the filter map. This map gets aggregated and weighed according to the decision made by the model. *ReLU* ensures that only those features are highlighted that had a positive effect on the decision making process.

Output of this step is an activation map that has dimensions equal to the dimensions of the convolution filter map. A bipolar interpolation of the activation map outputs the localization map.

7.1.2 Results. Using Grad-CAM we obtain the localization maps that highlight regions of importance on the input side. In figure 3, we see that the upper portion of the man's body is highlighted more compared to his lower part. This signifies that for making a decision, the model focused more on the head portion, which is anomalous as it is covered with a hood.

8 CONCLUSION

We described an autoencoder model augmented with an analysis step called Grad-CAM. Two settings were presented; unsupervised and semi-supervised with user ratings. For inference, we discussed thresholding, sorting and anomaly summarizers. Grad-CAM provided insight into the inner working of the model.

In future, instead of 2 ratings, we can use a hierarchy of ratings. Higher the rating, more is the normalcy associated with the frame. Also, we can include knowledge of objects explicitly. That way, the model will learn about the potentially dangerous and suspicious objects like gun or knife.

REFERENCES

[1] CUHK campus. 2013. Avenue Dataset for Abnormal Event Detection. (2013).

[2] Yong Shean Chong and Yong Haur Tay. 2017. *Abnormal Event Detection in Videos Using Spatiotemporal Autoencoder*. Springer International Publishing, 189–196. https://doi.org/10.1007/978-3-319-59081-3_23

[3] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe. 2015. Learning deep representations of appearance and motion for anomalous event detection. *British Machine Vision Conference, 2015* (2015).

[4] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury and L. S. Davis. 2016. Learning temporal regularity in video sequences. *Computer Vision and Pattern Recognition, 2016* (2016).

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Conference on Computer Vision* (2017), 618–626.

[6] T. Hospedales, S. Gong, and T. Xiang. 2009. A markov clustering topic model for mining behaviour in video. *International Conference on Computer Vision, 2009* (2009).

[7] V. Saligrama Y. Benezeth, P. Jodoin and C. Rosenberger. 2009. Abnormal events detection based on spatio-temporal co-occurrences. *in Proc. IEEE Conf. Comput. Vision Pattern Recog. Workshops, Jun.2025, 2009* (2009), 2458–2465.

[8] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas. 2011. Abnormal detection using interaction energy potentials. *Computer Vision and Pattern Recognition, 2011* (2011).

[9] C. Shen Y. Liu H. Lu Y. Zhao, B. Deng and X. Hua. [n. d.]. Spatio-Temporal AutoEncoder for Video Anomaly Detection. ([n. d.]), 1933–1941. https://doi.org/10.1145/3123266.3123451

[10] C. Zhou and R. C. Paffenroth. [n. d.]. Anomaly Detection with Robust Deep Autoencoders. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* ([n. d.]), 665–674.