Smart Surveillance

Abhishek Kumar
Dept. of Computer Science and Engineering
Indian Institute of Technology, Kanpur
abhikmr@cse.iitk.ac.in

Manish Kumar Bera
Dept. of Computer Science and Engineering
Indian Institute of Technology, Kanpur
mkbera@cse.iitk.ac.in

This is the project report for UGP(CS396) in Spring 2018 under Dr. Medha Atre.

Abstract

Modern day surveillance systems are mostly composed of CCTV cameras which need a human operator whose job is basically to sit down and look at the footage and wait for some anomalous event. However such a system is clearly inefficient because human operator will suffer from cognitive overload very soon. Also, these days we have a lot more CCTV cameras than the human operators. We try to come up with automated methods for surveillance. We use recent developments in NLP to learn basic definition of anomaly and integrate the Tensorflow's Object Detection API with the concepts of Information Theory to detect anomalous events in videos. A highlight of our project is that we remain context agnostic, i.e. our method will work effectively in any context and our method is unsupervised

Introduction

Surveillance systems these days consists of CCTV cameras which record the footage and a human operator keeps looking at it and wait for some anomalous event. As we can imagine that the operator will very soon suffer from cognitive overload. Therefore the efficiency of real time surveillance will decrease. Also with so many CCTV cameras these days, hiring many operator is an economic burden. There are about 4.2 million CCTV camera in Britain [1]. Clearly we can not imagine to have human operators to look at all of these and detect event in real time. Sometimes CCTV footage is recorded and later watched to find anomalous events. Such a system is clearly ineffective. Ideally we will want real time surveillance.

Automatic methods can solve these problems by increasing efficiency of video monitoring process. However research on real time fully automated smart surveillance is very limited. Also video is much richer data in the sense that it has 30frames per second thus is very difficult to process in real time. We try to come up with a method that can detect anomalous events by learning the context over a period of time. We try to keep ourselves context agnostic while developing this method. Our method can be deployed at any place with minimal cost requirements and needs no human intervention thereafter. Rest of the paper is organized in following way. First, we briefly discuss about some existing work in this area. Then we present our method and some some experimental results. Finally, we conclude with future scope.

Related Work

Medel and Savkis [2] learn generative models that can identify anomalies in videos using limited supervision. They train a composite Conv-LSTM that predict evolution of video sequence from a small number of input frames. Varghese et al. [3] propose a supervised algorithm which can work in very confined areas. Their algorithm uses statistical functions to learn motion path and speed of objects in video during training time. Chen et al.[4] develop a probabilistic framework to account for local-spatio temporal anomalies. Their approach is data driven. Cheng et al. [5] use hierarchical framework for detecting local and global anomalies. Local anomalies is detected using 3D pattern matching problem while global ones are detected using Gaussian process regression by extracting spatio-temporal features. Li et al. [6] learn normal trajectories from training videos and during test time report anomaly when observed trajectory is different from learnt. Blair et al. [7] develop a framework for anomaly detection which can work on FPGA, CPU and GPU. They focus on making it power aware i.e. based on recent frames it decides whether to process next frame on GPU or CPU or FPGA. They use Histogram of Oriented Gradients(HOG) for detecting cars and people and Mixture of Gaussians for detecting motions. Leyva et al. [8] come up with compact set of highly descriptive features for videos and identify the key regions. They use methods of optical flow and foreground occupancy to extract out a limited feature set. They try out Gaussian Mixture Models, Markov Chains and Bag Of Words on the extracted feature set. Achim et al. [1] use foreground object tracking and statistical scene modeller based on Bayesian theory to report anomaly. Their approach is unsupervised.

A common problem with all these approaches is that they are either heavily vision based and require lots of training data and time or they are heavily dependent on context. We

try to address these issues.

Our Method

NLP

We first try to get Abstractive summarization of video based on objects detected in video. For this we use NLP. We try to identify objects can be referred as inherently anomalous in urban settings. We wanted to do this by doing clustering over word vector representation and identify clusters which can represent anomalous words/objects/events. However defining what is anomalous is completely context based. We needed to manually curate some words that we can identify as anomalous. However, this method could have induced our personal bias. To eliminate this, we went over oxford dictionary once manually and made a list of words that can be anomalous in urban settings. At this stage we didn't worry about whether the word is detectable(i.e. it is an object) or not in videos. For e.g. we will include even anomalous verbs as well. It helps because we might have left some other object which is related to this word and is anomalous. For e.g. "hitting" might be related to "stick".

To learn other anomalous words we used three things. Firstly we computed $k-$nearest neighbours of the words from **Word2Vec** embeddings of these words [9]. Next we computed $k-$nearest neighbours of the words from **GloVe** embeddings [10]. We also picked $k-$synonyms of these words from their synsets in **WordNet**[1]. At the end of this step we have a comprehensive list of all the words(including nouns, verbs, adjectives) that may be considered inherently anomalous.

Next task was to refine the seeds. We needed to separate out the detectable objects from the list obtained in previous part. We first used *nltk* based tokenizer for this task but it wasn't giving good results. We realized that *nltk* based tokenizer was trained to work in sentences where context becomes very important. For eg.
1. Running is good for health.
2. The robber was running.
In sentence 1 "Running" is noun while in second sentence "running" is verb. We simply had a list of words that we wanted to classify as objects or non objects. So we manually went over the list and separated out detectable anomalous objects. Some of them are present in existing Tensor flow library while some aren't.

Feature Extraction

To extract features from video, we have used for simple methods right now. This is because right now we wanted to just test the working of information theoretic algorithm(explained in next section). However, one of the key highlights of our work is plug-and-play architecture. Later someone can very easily put in more complex, rich and less abstractive feature in place of our current features. For each frame we define a feature vector in $\mathbb{R}^3$. First component is number of objects detected in frame. The intuition is that if there are too many objects in single frame then the frame may be anomalous. Second is sum of maximum similarity of all the objects detected in frame with the list of inherently anomalous objects prepared in previous part. Third is average difference of pixel value of this frame with previous frame. This component will help us model history and learn small amount of context in online fashion. To achieve real time processing we needed to drop few frames to maintain the speed. We prepare feature vector of each processed frame and try to predict anomaly in next section.

Semantic Search

This is one of the key highlights of our project. Given an object we output the times at which this object or an object with same semantic meaning was found in a long video sequence. For eg. if one searches for cat and there was no cat in video but a dog or some other pet at few places in video, then we report those frames/time. This eliminates the need to go over the entire video sequence manually. This can be very useful if someone wants to do analysis of long video sequences.

Information Theory

Information theory has been used to detect anomaly in *sendmail* data[11]. We will also use it to detect anomaly in online fashion. First we review some basics of information theory.

(Shannon & Weaver [12]) Definiton 1 *For a dataset X where each data item belongs to a class $x \in C_x$, the entropy of X relative to this $|C_x|$ wise classification is defined as $H(X) = -\sum_{x \in C_x} P(x) \log P(x)$ where $P(x)$ is probability of x in X*

Definiton 2 *The conditional entropy of X given Y is the entropy of conditional distribution $P(x|y)$, that is $H(X|Y) = -\sum_{x \in C_x, u \in C_y} P(x,y) \log P(x|y)$ where $P(x|y)$ is conditional probability of x given y and $P(x,y)$ is joint probability distribution of x, y*

Definition 3 *The relative entropy between two distributions $p(x)$ and $q(x)$ over same $x \in C_x$ is defined as $relEntropy(p,q) = -\sum_{x \in C_x} p(x) \log \frac{q(x)}{p(x)}$*

Definition 4 *The relative conditional entropy between two distributions $p(x|y)$ and $q(x|y)$ that are defined over same $x \in C_x$ and $y \in C_y$ is $relCondEntropy = -\sum_{x \in C_x, y \in C_y} p(x,y) \log \frac{q(x|y)}{p(x|y)}$*

[1] `https://wordnet.princeton.edu/`

Now we will present our algorithm. Let X denote the sequence of observations $(e_1, e_2, e_3.....e_n)$ and $Y = (e_1, e_2, e_3, ...e_k)$ where $k < n$. Then conditional entropy $H(x|y)$ tells how much uncertainty is present in rest events in x given we have already observed events in y. Let's fix $k = n - 1$. In our case, let e_i denote feature vector representation of frame i. Suppose we have observed frame $e_1, e_2, ..e_{n-1}$ and we want to predict e_n. However $e_n \in R^3$. Let's say we discretize the domain of e_n and instead of predicting exact e_n we predict the class of e_n in the new discretized domain. Let l denote sequence length that we use while predicting e_n i.e. $l = n - 1$. Then intuitively misclassification rate should decrease as we increase l. Also, by intuition $H(X|Y)$ should also decrease when we increase l. Lee and Xiang prove our intuition[11] on *sendmail* data. Also misclassification rate goes hand in hand with conditional entropy (both decrease with sequence length and then saturate. We leverage this relationship to find out optimal sequence length. Note that calculating misclassification rate is computationally costly, so we use conditionall entropy to find the sequence length at which it saturates and use it as optimal sequnce length.

After fixing sequence length we now train our model for first few observations (say 1000). This is burn in period. For now we are keeping our model to be very small LSTM, however in future we wish to experiment with Hidden Markov Model and State Space Model. After training we keep predicting new points. We give a score based on difference in our predicted value and observed value. A high score will indicate presence of anomalous event.

Some times it may happen that our misclassification rate is poor and we are constantly giving high score. If misclassification rate and conditional entropy vary too much then it means that our model is not good and we need to update our model. Then we stop predicting and start a burn in phase where we use observed values for training. If relative conditional entropy between test and train is very high then it means that the sequence length we are using is not sufficient, so we increase the sequence length and retrain.

Experiments

We ran our tool on the following CC-TV videos (courtesy Dept. of CSE, IITK)(These CC-TV videos are of cameras of the CSE building):

1. Main door
2. Main door (inside view)
3. Corridor

Below we present output of our tool. Blue curve represents observed data points while orange denotes predicted.

- **sim score** stands for similarity with identified anomalous objects

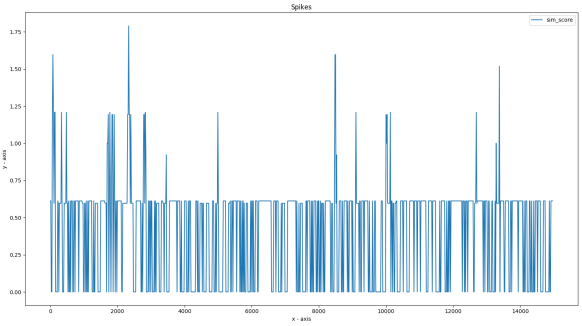

Figure 1. Object Detection : sim score for Main door morning:

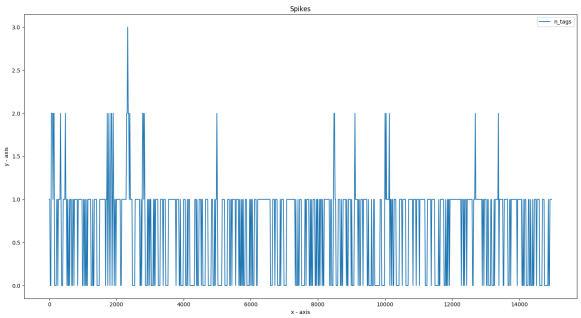

Figure 2. Object Detection : n tags for Main door morning:

- **n tags** stands for number of objects in frame
- **pixel difference** stands for average squared pixel difference of two images

Results from corridor have not been shown due to space constraints.

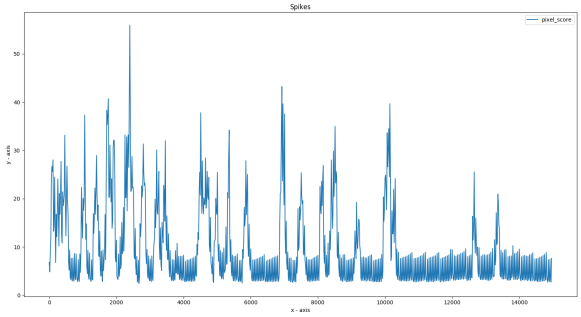

Figure 3. Object Detection : pixel score for Main door morning:

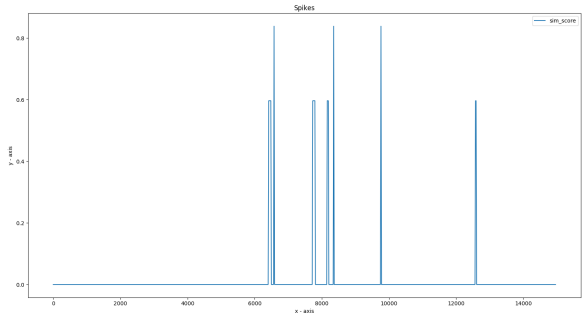

Figure 4. Object Detection : sim score for Main door midnight:

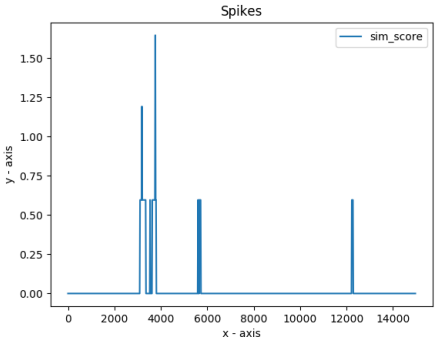

Figure 7. Object Detection : sim score for Main door(inside view) morning:

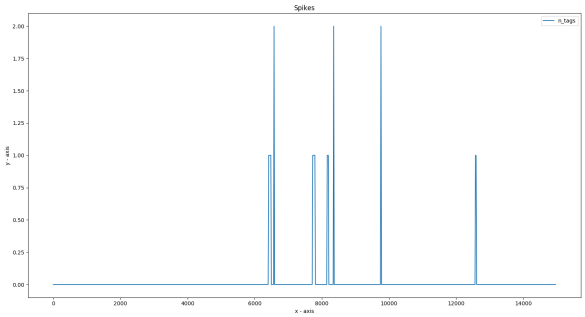

Figure 5. Object Detection : n tags for Main door midnight:

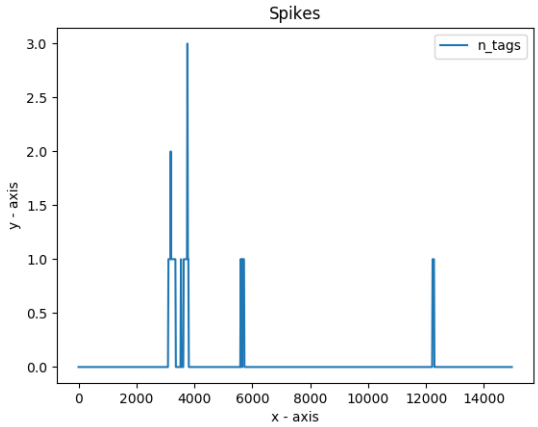

Figure 8. Object Detection : n tags for Main door (inside view) morning:

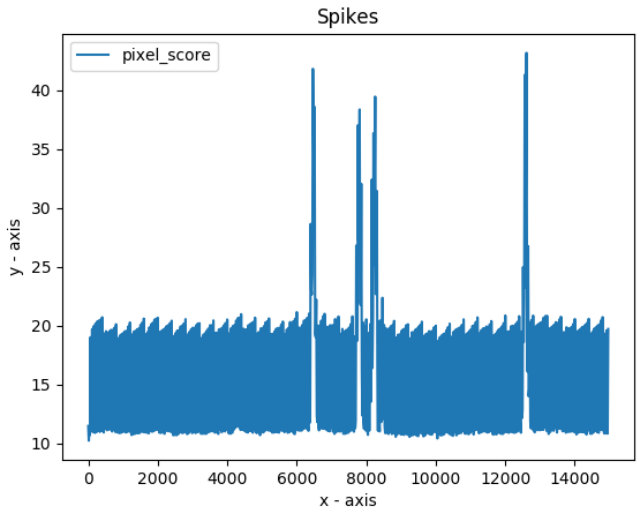

Figure 6. Object Detection : pixel score for Main door midnight:

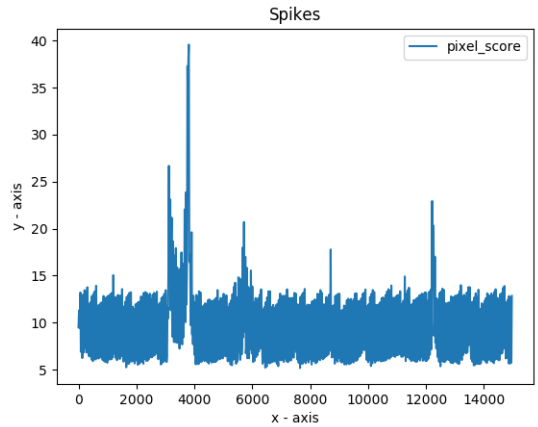

Figure 9. Object Detection : pixel score for Main door(inside view) morning:

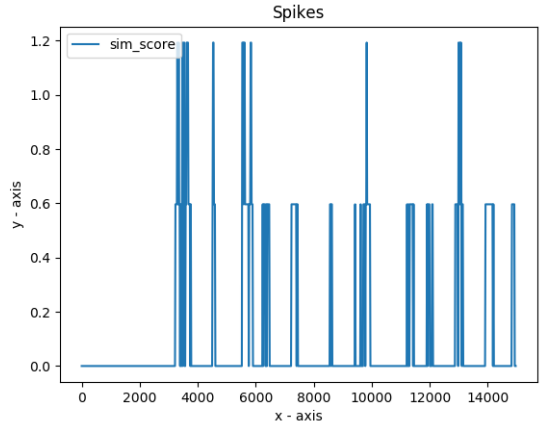

Figure 10. Object Detection : sim score for Main door(inside view) midnight:

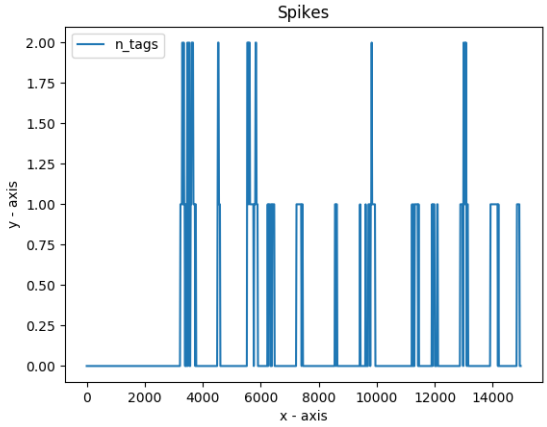

Figure 11. Object Detection : n tags for Main door (inside view) midnight:

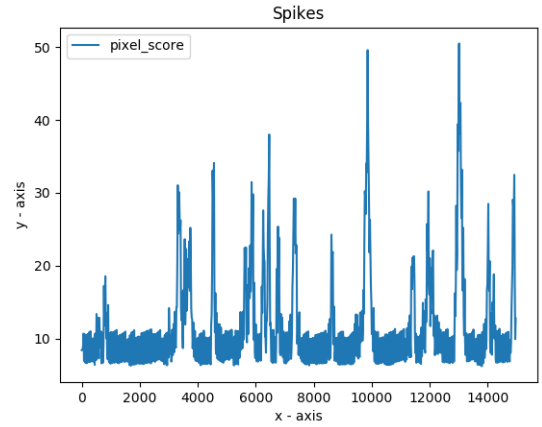

Figure 12. Object Detection : pixel score for Main door(inside view) midnight:

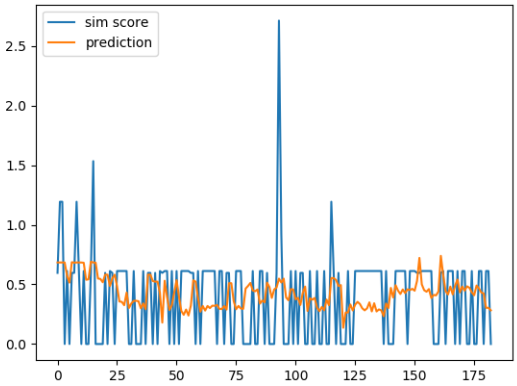

Figure 13. Anomaly detection : Main door morning:

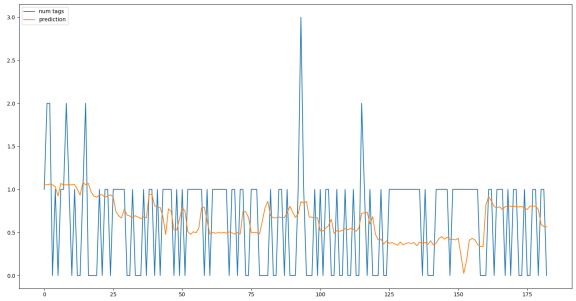

Figure 14. : Anomaly detection Main door morning:

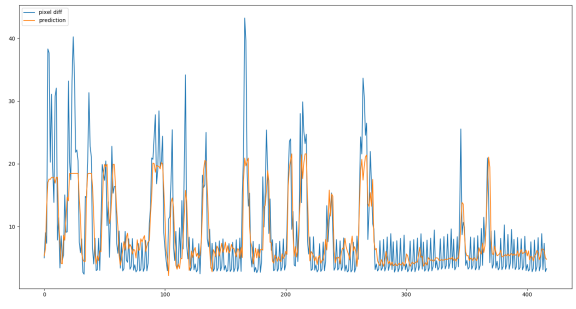

Figure 15. Anomaly detection Main door morning:

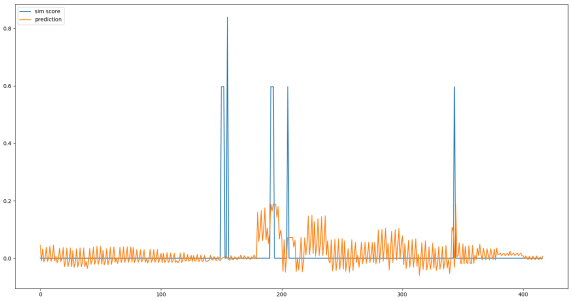

Figure 16. Anomaly detection Main door midnight:

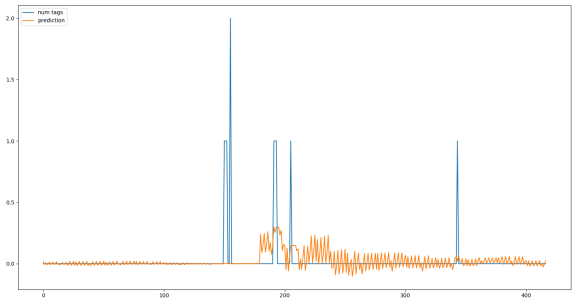

Figure 17. Main door midnight:

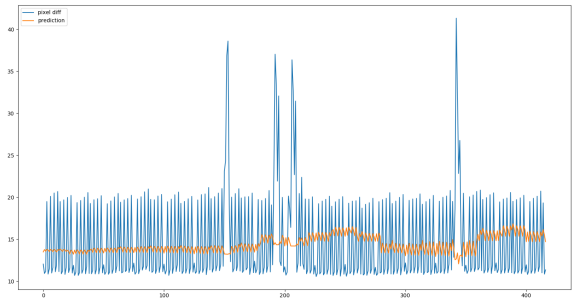

Figure 18. Main door midnight:

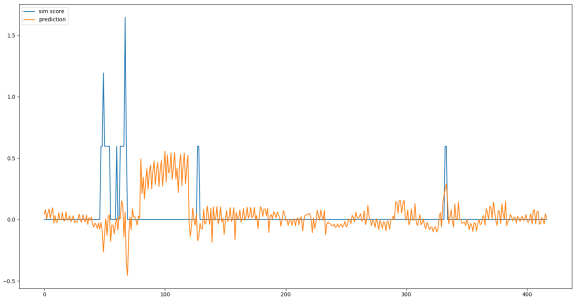

Figure 19. Main door (inside view) morning :

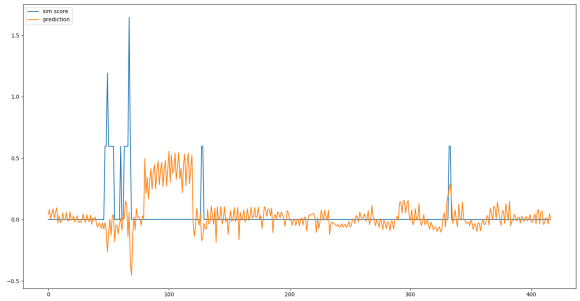

Figure 20. Main door (inside view) morning :

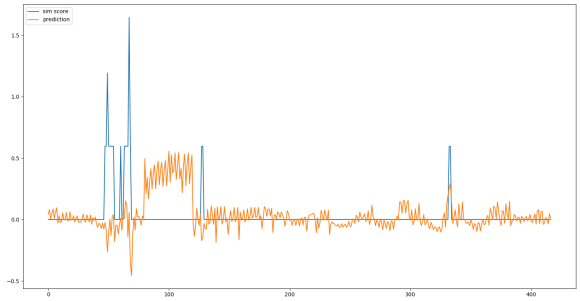

Figure 21. Main door (inside view) morning :

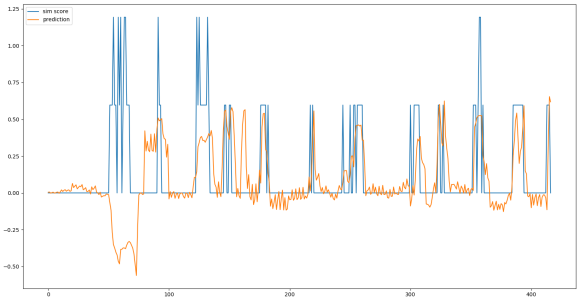

Figure 22. Main door (inside view) midnight :

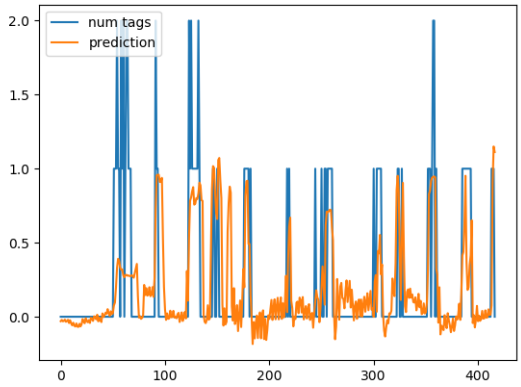

Figure 23. Main door (inside view) midnight :

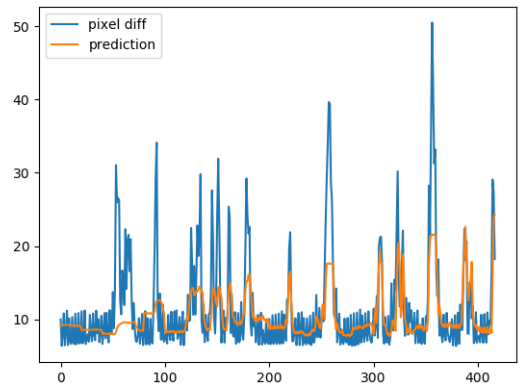

Figure 24. Main door (inside view) midnight :

Conclusion

We use concepts of information theory to come up with an unsupervised algorithm to identify anomaly in videos in context agnostic manner. A key highlight of our algorithm is that it is fully automated and needs no human intervention at all. It automatically decides sequence length as and when needed, adapts to the context and reports anomaly if observed frame is significantly different from what is expected from context. Our algorithm can be deployed for surveillance. For future work we want to change our current methodology of feature extraction and obtain much richer features. Also instead of using LSTM, we wish to use State Space Model and Hidden Markov Model for prediction. We also wish to incorporate coordination between multiple camera feeds and jointly detect anomalous event. Automated Surveillance is a very promising field considering the amount of cost incurred in hiring human operators and inefficiency due to human error.

References

[1] H. Li, A. Achim, and D. Bull, "Unsupervised video anomaly detection using feature clustering," *IET Signal Processing*, vol. 6, pp. 521–533, July 2012.

[2] J. R. Medel and A. E. Savakis, "Anomaly detection in video using predictive convolutional long short-term memory networks," *CoRR*, vol. abs/1612.00390, 2016.

[3] E. Varghese, J. Mulerikkal, and A. Mathew, "Video anomaly detection in confined areas," *Procedia Computer Science*, vol. 115, pp. 448 – 459, 2017. 7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India.

[4] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2112–2119, June 2012.

[5] K. W. Cheng, Y. T. Chen, and W. H. Fang, "Video anomaly detection and localization using hierarchical feature representation and gaussian process regression," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2909–2917, June 2015.

[6] X. Li and Z. m. Cai, "Anomaly detection techniques in surveillance videos," in *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pp. 54–59, Oct 2016.

[7] C. G. Blair and N. M. Robertson, "Video anomaly detection in real time on a power-aware heterogeneous platform," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, pp. 2109–2122, Nov 2016.

[8] R. Leyva, V. Sanchez, and C. T. Li, "Video anomaly detection with compact feature sets for online performance," *IEEE Transactions on Image Processing*, vol. 26, pp. 3463–3478, July 2017.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[10] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

[11] W. Lee and D. Xiang, "Information-theoretic measures for anomaly detection," in *Proceedings 2001 IEEE Symposium on Security and Privacy. S P 2001*, pp. 130–143, 2001.

[12] C. E. Shannon and W. Weaver, "The matliematical theory of communication," 1949.