Anomaly Detection

Dinesh Nariani **AU1920128**

Henil Shah **AU1940205**

Devanshu Magiawala **AU1940190**

Abstract — The detection of abnormal occurrences is critical in video content analysis and is a difficult task when monitoring surveillance fields. Surveillance cameras can capture a wide range of actual irregularities. The paper highlights a learning algorithm that trains on both normal and anomalous videos to learn to detect anomalies. To avoid spending time annotating anomalous segments or clips in training films, the proposed learning anomaly uses the deep multiple instance ranking framework and weakly labeled training movies, where the training labels (anomalous or normal) are at the video level rather than the clip level. In our approach, we use multiple instance learning (MIL) to automatically develop a deep anomaly ranking model that predicts high anomaly scores for anomalous video segments by treating normal and anomalous movies as bags and video segments as instances.

Keywords — Anomaly event detection, Multiple instance learning Framework, C3d Model, I3D Model, 3D convolutions, activity recognition, feature extraction, video surveillance, computer vision.

I. Introduction

The use of cameras for surveillance has been increasing with time at public places to ensure safety. One of the main tasks for video surveillance is the detection of anomalies such as crimes, illegal activities, fire etc. These types of anomalous events rarely occur as compared with normal activities. So, in order to slice off the waste of time and labor, developing and deploying an intelligent computer vision algorithm for anomaly detection through video surveillance is in great need. So, the primary target for these anomaly detection systems is to identify and detect anomalies and generate a signal which is different from the signal generated through normal patterns. All these real world anomalies happen for a short duration and are complicated, due to which it is not possible to list down all possible anomalous events. So, it is necessary and appropriate that anomaly detection algorithms do not depend on any prior information about the events. Also, in these types of algorithms, as the situation captured by the cameras can change over time, these could lead to high false positives for different normal behaviors.

II. Literature Survey

Abnormal detection is one of the most difficult and time-consuming tasks in computer vision. A number of attempts have been made in surveillance systems to detect abnormal activity.[1] Yihao Zhang et al.proposes anomaly detection in traffic video with the information provided in the HEVC compressed domain.In High Efficiency Video Coding (HEVC).[2] Tian Wanga et al. propose event detection based on moment feature descriptor and classification. The feature descriptor extracts the optical flow and computes the histogram of optical flow orientations(HOFO).The hidden Markov model(HMM) is proposed to classify the events due to the probabilistic property. HOFO feature descriptor is based on the movement and if the index of the HOFO is at a low level.[3]

Fan Jiang et al. propose a multi-sample-based similarity measure in which HMM training and distance measurement are performed on many samples. A unique dynamic hierarchical clustering (DHC) algorithm is used to collect these numerous

training data sets.[4] Xun Tanget et al. offer a method for recognizing anomalous occurrences in crowded settings based on sparsely coded motion attention.[5]

III. Dataset

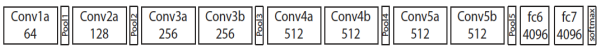
The dataset used for the proposed approach is the anomaly detection dataset from the University of Central Florida (UCF). The dataset contains a total of 1692 surveillance films, with 900 anomalous videos and 792 normal videos. Arrest, Arson, Assault, Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism are among the 13 real-world abnormalities and anomalous incidents covered in the videos. We are considering 5 videos for each anomaly, 65 videos for normal videos, and 10 videos from each abnormality for testing purposes for our project. We are using multiple instances learning to annotate our dataset. We are considering this particular sample for training our i3d model. We consider another sample which consists of 130 anomaly videos and 130 normal videos. Each anomaly has 10 videos. This sample will be trained again on a C3D model.

IV. Multiple Instance Learning

MIL eliminates the requirement for correct temporal annotations. The specific temporal position of anomalous events in videos is uncertain in MIL. Only video-level labels identifying the presence of an anomaly in the entire video are required. A video with anomalies is labeled as positive, while a video with no anomalies is labeled as negative. Then, we represent a positive video as a positive bag B_a, with different temporal segments forming individual instances in the bag (p^1, p^2,..., p^m), where m is the number of instances in the bag. We presume that the anomaly exists in at least one of these circumstances. Similarly, the negative video is represented by a negative bag B_n, in which temporal segments generate negative instances (n^1, n^2, . . . , n^m).

V. C3D Model

C3D convolutional neural networks are deep 3-dimensional convolutional neural networks with a homogeneous design consisting of 3 x 3 x 3 convolutional kernels followed by 2 x 2 x 2 pooling at each layer. They've been trained on large-scale supervised video datasets including UCF-101 and Sports 1M.

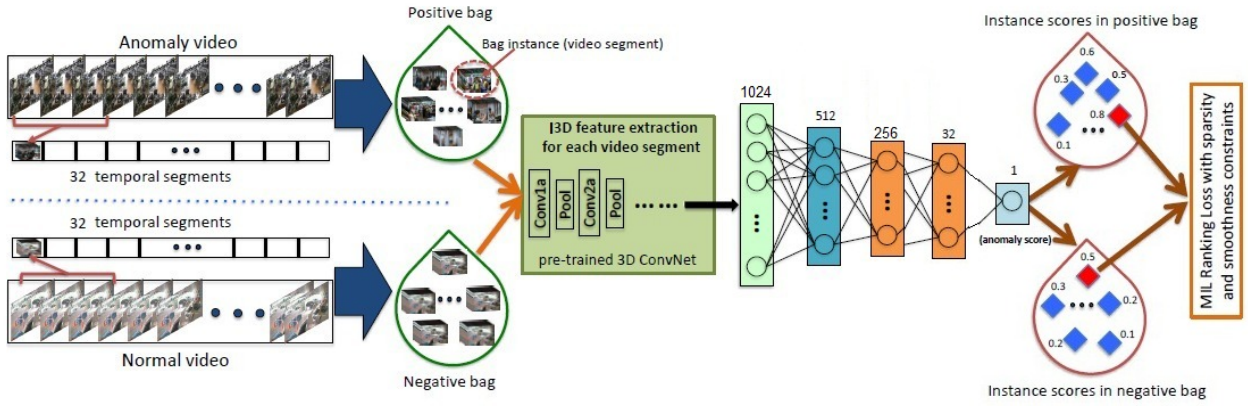


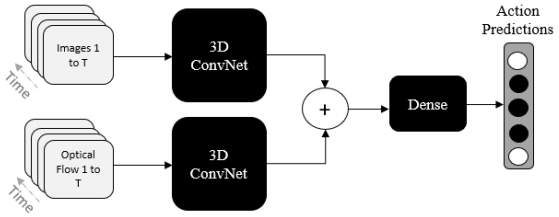
The first convolution layer of size 1x3x3 is depicted in the figure, followed by a pooling layer of size 1x2x2. This is done to maintain temporal information in the first layer and generate higher level temporal representations in later levels of the network. Every other convolution layer and pooling layer would be 3x3x3 and 2x2x2, respectively, with strides of 1 and 2. The fully connected layers have 4096 dimensions and softmax outputs that represent either 101 classes from the UCF-101 dataset or 487 classes from the Sports 1M dataset. For our project, we will look at the layer fc6 output, which is a 4096 feature vector with 16 frames.

VI. I3D Model

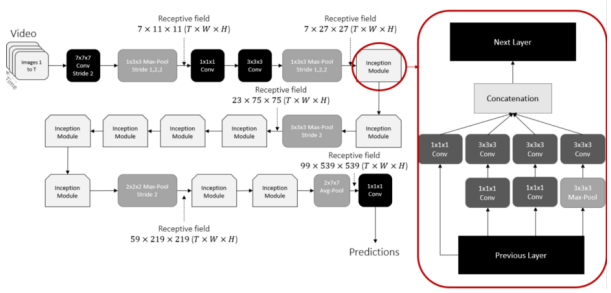
To compare the results from the C3D model, another model named I3D model is also used. The approach works by inflating the 2D architecture of all the filters and pooling kernels. As we know, filters in the 2D model are square N x N, now inflating them turns them into a cube of size N x N x N.

I3D Convolution applies a bootstrap 3D filter from 2D filters on already trained models. Here, by constantly replicating a picture into a video sequence, an image can be transformed into a video. This can be accomplished by repeating the weights of the 2D filters N times along the time dimension and rescaling them by dividing by N, owing to linearity. The outputs of pointwise nonlinearity layers, as well as average and max-pooling layers, are identical to those of the 2D case.





In a convolutional neural network, a receptive field is the fraction of the image that is exposed to one filter at a time, and it grows as we stack more layers. 2D convolutions and pooling are symmetrical since they focus on the image's height and width. If the receptive field expands too quickly in time compared to space, edges from various objects may become conflated, preventing early feature recognition. It's possible that if the receptive field expands too slowly, it won't be able to catch scene dynamics as well. In conclusion, due to the added time dimension, the kernels in I3D are not symmetrical.



The beginning of the network, as shown in the diagram, uses asymmetrical filters for max-pooling, keeping time while pooling over the spatial dimension. It doesn't execute convolutions and pooling that include the time dimension until later in the network. Overall, it's a close approximation to an ideal local sparse structure. It also collects the findings after processing spatial (and time in this example) data at various scales. The goal of this module was to allow the network to become "wider" rather than "deeper."

VII. Implementation

We extract visual features from the C3D network's fully connected (FC) layer FC6. Before we compute features, we resize each video frame to 112 x 112 pixels and set the frame rate to 30 frames per second. After l2 normalization, we compute C3D characteristics for each 16-frame video clip. To get features for a video segment, we average all 16-frame clip features inside that segment. These characteristics (4096D) were fed into a 3-layer FC neural network. The first FC layer comprises 512 units, then 32 units, and finally one unit. Between FC layers, dropout regularization of 60% is utilized. We use ReLU activation and Sigmoid activation for the first and the last FC layers respectively, and employ an Adagrad optimizer with the initial learning rate of

0.001. The parameters of sparsity and smoothness constraint in the MIL ranking loss are set to $\lambda_1=\lambda_2 = 8\times10^{-5}$ and $\lambda_3 = 0.01$ for the best performance.This approach will be applied to the new sample of 260 videos.

We extract visual features from the I3D network's from the global average pooling. Before we compute features, we resize each video frame to 224 x 224 pixels. We compute I3D characteristics for each 16-frame video clip. To get features for a video segment, we average all 16-frame clip features inside that segment. These characteristics (1024D) were fed into a 4-layer FC neural network. The first FC layer comprises 512 units, then 256 units and then 32 units, and finally one unit. We use ReLU activation and Sigmoid activation for the first and the last FC layers respectively, and employ an Adagrad optimizer with the initial learning rate of 0.001. The parameters of sparsity and smoothness constraint in the MIL ranking loss are set to $\lambda_1=\lambda_2 = 8\times10^{-5}$ and $\lambda_3 = 0.01$ for the best performance. This approach will be applied to the new sample of 130 videos. This would be compared with the previously trained C3D model with the sample of 130 videos.

VIII. Loss function

This is a regression problem, according to us. As a result, our ranking loss in the hinge-loss formulation is as follows:

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i\in\mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i\in\mathcal{B}_n} f(\mathcal{V}_n^i)).$$

One limitation of the above loss is that it ignores the abnormal video's underlying temporal structure. First, in real-world circumstances, abnormality frequently occurs for a short period of time. In this circumstance, the scores of the instances (segments) in the anomalous bag should be sparse, indicating that the anomaly may be contained in only a few segments. Second, because the video is a series of segments, the anomaly score should vary gradually between them. As a result, by minimizing the difference in scores for adjacent video segments, we impose temporal smoothness between anomaly scores of temporally adjacent video segments. The loss function is transformed by including the sparsity and smoothness restrictions on the instance scores.

$$l(\mathcal{B}_a, \mathcal{B}_n) = \max(0, 1 - \max_{i\in\mathcal{B}_a} f(\mathcal{V}_a^i) + \max_{i\in\mathcal{B}_n} f(\mathcal{V}_n^i))$$

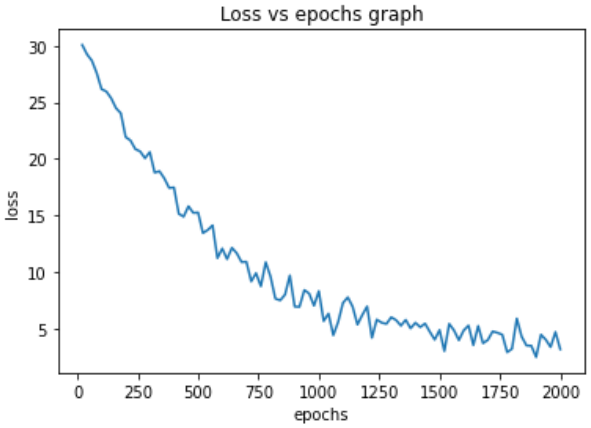
$$+\lambda_1 \overbrace{\sum_{i}^{(n-1)} (f(\mathcal{V}_a^i) - f(\mathcal{V}_a^{i+1}))^2}^{①} + \lambda_2 \overbrace{\sum_{i}^{n} f(\mathcal{V}_a^i)}^{②},$$

where term 1 indicates temporal smoothness and term 2 indicates sparsity.

IX. Results

A. C3D results on 130 Video Sample:

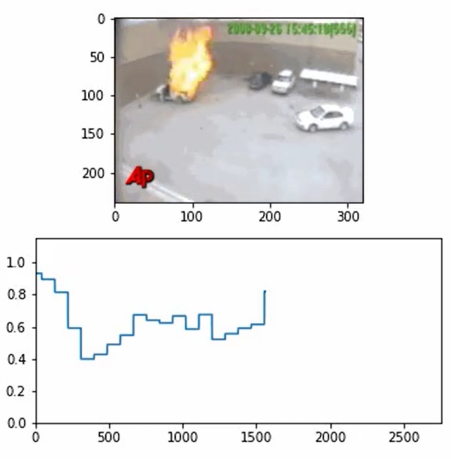
We have trained our model for 2000 iterations, batch size is 60, learning rate is 0.01 and we have got the sum of hinge-loss, sparsity loss and smoothness loss which is 5.38.



We generated some results on testing videos. Due to consideration of small sample size, we got some false negative results in anomalies while testing manually like abuse, arrest, assault, road accidents, robbery, shooting, shoplifting.

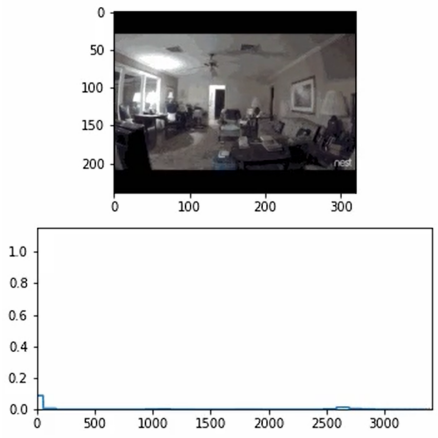
True Positive: -

Explosion Anomaly



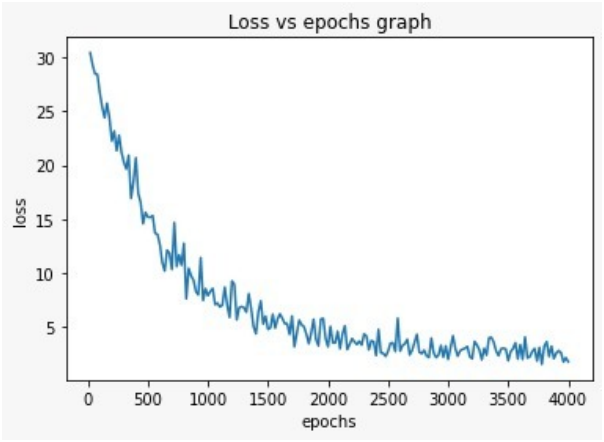
False Negative: -

Abuse Anomaly



B. C3D results on 260 Video Sample:

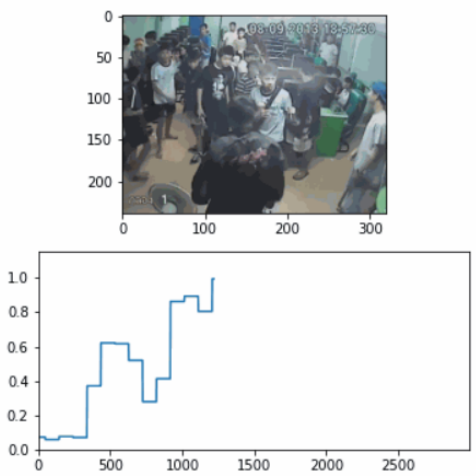
We have trained our model for 4000 iterations, batch size is 32, learning rate is 0.01 and we have got the sum of hinge-loss, sparsity loss and smoothness loss which is 1.7413.



We generated some results on testing videos. We increased the size of the sample, we got some improvements from our previous approach and negative results in anomalies. We are able to get true positives in abuse, arrest, arson, assault, fighting, robbery, shoplifting, vandalism, and shooting, stealing. Still, there were false negatives in burglary, explosion, road accidents. While we are able to generate decent results with other anomalies.It is important to note that, despite the absence of segment-level annotations, the network can estimate the temporal position of an anomaly based on anomaly scores.As the iterations increases and the network sees more videos of the dataset, it will automatically learns to precisely localize anomaly.

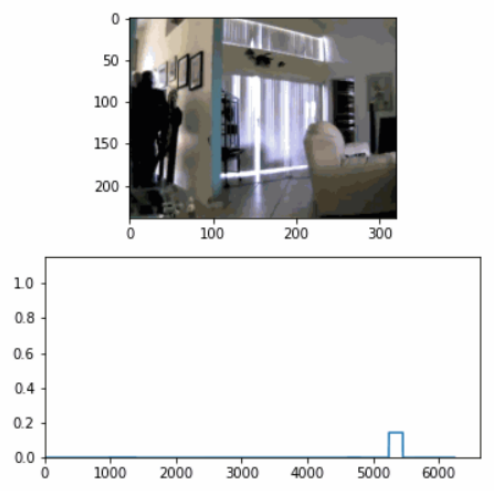
True Positive: -

Fighting Anomaly



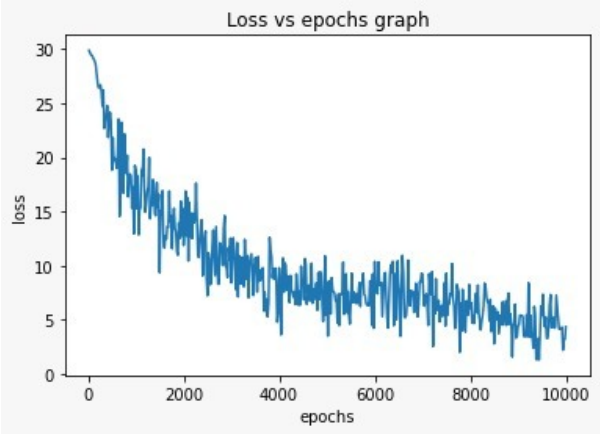
False Negative: -

Burglary Anomaly



C. I3D results on 130 Video Sample:

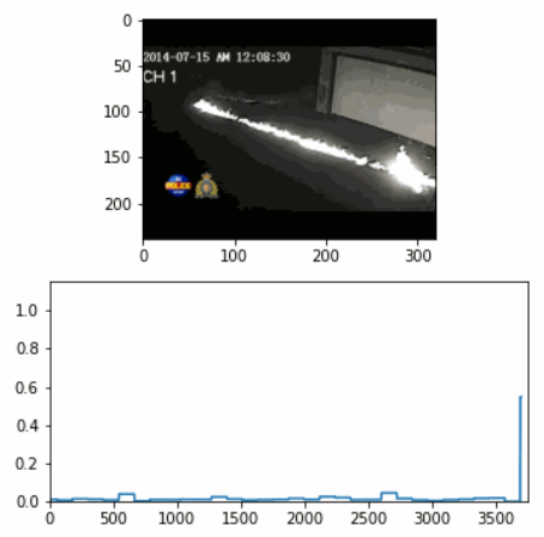
We have trained our I3d model for 10000 iterations, batch size is 32, learning rate is 0.01 and we have got the sum of hinge-loss, sparsity loss and smoothness loss which is 2.23.



We generated some results on testing videos. Due to consideration of small sample size, we got some false negative results in anomalies while testing manually like abuse, arrest, assault, road accidents, robbery, shooting, shoplifting. While we are able to generate decent results with other anomalies.It is important to note that, despite the absence of segment-level annotations, the network can estimate the temporal position of an anomaly based on anomaly scores. As the iterations increase and the network sees more videos of the dataset, it will automatically learn to precisely localize anomalies.

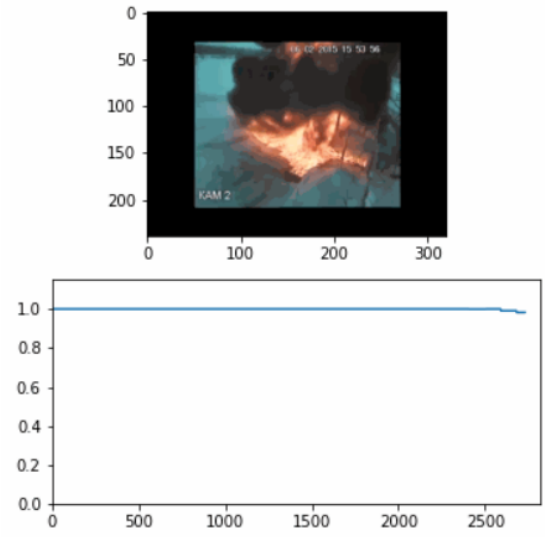
True Positive: -

Arson Anomaly



False Negative: -

Explosion Anomaly



X. Conclusion

Based on the models and approaches used, for the firstly trained C3D model considering 5 videos of each of 13 anomalies, the model was able to predict 5 different kinds of anomalies positively, while there were 8 false negative anomalies. So, by increasing the sample size where 10 videos for each anomalies were considered for training the model, the results came out to be 9 true positive anomalies detection and 4 false negatives. Also, another approach that was used was by switching the C3D model to I3D model and considering 5 videos for each of 13 anomalies, the results turned out that 6 true positive anomalies were recognized while there were still 7 false negatives. In the future works one can try on increasing the sample size and trying different features extractor to get best results.

XI. References

1. S. Mohammadi, A. Perina, H. Kiani, and M. Vittorio.Angry Crowds: Detecting violent events in videos. In ECCV, 2016.
2. Yihao Zhang, Hongyang Chao.Abnormal Event Detection in Surveillance Video: ACompressed Domain Approach for HEVC, 2017.
3. Tian Wanga, MeinaQiaoa, Yingjun Dengb, Yi Zhouc, HuarenWang, Qi Yuan, and Hichem Snoussi.Abnormal Event Detection based on Analysis of Movement Information of Video Sequence, 2017.
4. Fan Jiang, Ying Wu, Aggelos K. Katsaggelos.Abnormal Event Detection From Surveillance Video By Dynamic Hierarchical Clustering, 2007.
5. Xun Tang, Shengping Zhang, Hongxun Yao.Sparse Coding Based Motion Attention For Abnormal Event Detection,2013.
6. Sultani, W. and Shah, M., 2022. Real-World Anomaly Detection in Surveillance Videos. [online] Openaccess.thecvf.com. Available at: <https://openaccess.thecvf.com/content_cvpr_2018/html/Sultani_Real-World_Anomaly_Detection_CVPR_2018_paper.html> [Accessed 20 March 2022].
7. Medium. 2022. *Deep Dive into Convolutional 3D features for action and activity recognition (C3D)*. [online] Available at: <https://medium.com/@nair.binum/quick-overview-of-convolutional-3d-features-for-action-and-activity-recognition-c3d-138f96d58d8f> [Accessed 20 March 2022].
8. Medium. 2022. *Deep Dive into Convolutional 3D features for action and activity recognition (C3D)*. [online] Available at: <https://medium.com/@nair.binum/quick-overview-of-convolutional-3d-features-for-action-and-activity-recognition-c3d-138f96d58d8f> [Accessed 20 March 2022].
9. "Understanding the Backbone of Video Classification: The I3D Architecture", Medium, 2022. [Online]. Available: https://towardsdatascience.com/understanding-the-backbone-of-video-classification-the-i3d-architecture-d4011391692. [Accessed: 24- Apr- 2022].