**Anomaly Detection in Surveillance Videos**

**1. Introduction**

To improve public safety, surveillance cameras are being utilized more frequently in public spaces like roadways, junctions, banks, and shopping centers. Law enforcement organizations' monitoring capacities, however, have not kept up. As a result, there is a conspicuous deficit in the use of security cameras and an inefficient camera-to-monitor ratio. Identifying abnormal events, such as traffic accidents, crimes, or unlawful activity, is a crucial duty in video surveillance. As opposed to usual activity, anomalous events typically happen far less frequently. Therefore, there is an urgent need to create computer vision algorithms for automatic video anomaly identification to reduce labor and time loss. A practical anomaly detection system's objective is to promptly report behavior that deviates from expected patterns and pinpoint the time window in which the anomaly is occurring. As a result, anomaly detection can be thought of as a basic form of video understanding that separates anomalies from expected patterns.

**2.Executive summary:**

It is a problem of anomaly detection in surveillance videos. The workflow for this project starts from collecting anomaly and normal videos, processing and extracting the features from the videos. For example, let’s consider an anomaly video assume that we have three anomalies in that video, we initially start converting the video into frames, from the frames we apply featurization techniques say covolution3D or Inflated3D on it. We send the feature to training module where it has sequence of convolution, pooling, dropout and dense layer at the end in case of C3D or I3D. in the later stage we get some anomaly scores marked on each frame. In C3D we used 2 epochs with 20000 iterations to extract the features. Whereas in the case of I3D we used 7335 epochs. With C3D we got an AUC score of 0.68 but with I3D it is 0.962. here I3D is more efficient in extracting the features from the data and so the training is more efficient.so the accuracy is high comparatively.

**3. Problem Statement:**

Surveillance videos act as proof of evidence, it does not act as a live interacting system. What if a surveillance camera can act as a threat reporting system. The goal of the project is to implement a deep learning technique that can take advantage of surveillance videos and report any unusual activities. This technique is known as anomaly detection.

Once an anomaly is found, it can then be classified using classification algorithms into one of the specific tasks. Creating algorithms to identify a particular anomalous occurrence, such as a violence detector or a traffic accident detector, is a first step in tackling anomaly detection. However, such methods cannot be applied to identify further anomalous events, therefore in reality, their practical utility is quite restricted.

**4. Techniques**

A wide range of believable anomalies can be seen in surveillance videos. In this study, we suggest using both regular and anomalous films to learn abnormalities. We suggest learning anomaly through the deep multiple instances ranking framework by utilizing weakly labeled training videos, where the training labels (anomalous or normal) are at the video level rather than the clip-level, to avoid spending a lot of time annotating the anomalous segments or clips in training videos. Our method uses multiple instance learning (MIL), where video segments are instances and normal and abnormal movies are bags. This method automatically learns a deep anomaly ranking model that forecasts high anomaly scores for abnormal video segments. To locate anomaly more accurately during training, we also include sparsity and temporal smoothness restrictions in the ranking loss function. By requiring consistent optical flow between predicted frames and ground truth frames, we add a motion (temporal) constraint to video prediction in addition to the usual appearance (spatial) constraints on intensity and gradient. The video prediction task is the first to incorporate a temporal constraint in this work.

Used methods:

1. Methods based on hand-crafted features

2.Methods Using Deep Learning

The underlying assumption of the proposed method is that the network can automatically learn to predict the location of the anomaly in a video when presented with many positive and negative videos labeled at the video level. During training iterations, the network should acquire the ability to produce high scores for anomalous video segments to achieve this objective.

**Technique 1:**

This technique is to separate each of the positive (contain anomaly anywhere) and negative (contain no anomaly) videos into several temporal video chunks. Each temporal segment then represents an instance in the bag, which is represented by each video as a bag. We train a fully connected neural network using a unique ranking loss function after extracting C3D features for video segments. This function calculates the ranking loss between the top scored occurrences (shown in red) in the positive bag and the negative bag.

Diagram

Description automatically generated

Although current methods show effective detection performance, their recognition of the positive instances, i.e., rare abnormal snippets in the abnormal videos, is largely biased by the dominant negative instances, especially when the abnormal events are subtle anomalies that exhibit only small differences compared with normal events.

In standard supervised classification problems using support vector machine, the labels of all positive and negative examples are available, and the classifier is learned using the following optimization function

Diagram

Description automatically generated

where k is the total number of training examples, w is the classifier to be learnt, yi represents the label of each case, x symbolizes feature representation of an image patch or video segment, b is a bias, and 1 is the hinge loss. Accurate annotations of positive and negative instances are necessary to develop a robust classifier. A classifier needs temporal annotations for each video segment in the context of supervised anomaly detection. However, getting temporal annotations for videos is a tedious and time-consuming process. The requirement of providing these precise temporal annotations is relaxed by MIL. MIL does not know the exact temporal positions of abnormal events in videos. Instead, all that is required are video-level labels showing the presence of an abnormality throughout the entire video. Positive labels are applied to videos having abnormalities, and negative labels are applied to videos without any anomalies. Then, a positive video is represented as a positive bag Ba, where different temporal segments create distinct instances in the bag, (p1, p2,..., pm), where m is the total number of instances in the bag. We presume that the anomaly is present in at least one of these situations. Similar to this, a negative bag, Bn, is used to represent the negative video, where temporal segments generate negative instances (n1, n2,.. nm). None of the examples in the negative bag have an oddity. One can optimize the objective function since the precise information (i.e., instance-level label) of the positive examples is unknown

Text

Description automatically generated

where YBj denotes bag-level label, z is the total number of bags.

During training, this strategy begins by dividing surveillance videos into predetermined segments. A bag is made up of these segments. The proposed deep MIL ranking loss is used to train the anomaly detection model with positive (anomalous) and negative (normal) bags.

**Technique 2**:

RTFM methodology developed on top of MIL approach, and it is accepted at ICCV 2021. t trains a feature magnitude learning function to effectively recognize the positive instances, substantially improving the robustness of the MIL approach to the negative instances from abnormal videos. RTFM also adapts dilated convolutions and self-attention mechanisms to capture long- and short-range temporal dependencies to learn the feature magnitude more faithfully.

Diagram

Description automatically generated

RTFM receives a T × D feature matrix F extracted from a video containing T snippets. Then, MTN captures the long and short-range temporal dependencies between snippet features to produce X = sθ(F). Next, we maximize the separability between abnormal and normal video features and train a snippet classifier using the top-k largest magnitude feature snippets from abnormal and normal videos.

**5. Conclusions**

We use the frame-level area under the ROC curve (AUC) as the evaluation measure for all data sets. Moreover, following we also use average precision (AP) as the evaluation measure Larger AUC and AP values indicate better performance. Some recent studies recommend using the region-based detection criterion (RBDC) and the track-based detection criterion (TBDC) to complement the AUC measure, but these two measures are inapplicable in the weakly supervised setting. Thus, we focus on the AUC and AP measures

**Results using C3D Feature extraction**:

**ROC \_AUC**:

Chart, line chart

Description automatically generated

Training images for 2 epochs:

A screenshot of a computer

Description automatically generated with medium confidence

Loss Images:

Graphical user interface, application

Description automatically generated

**Results using I3D Feature extraction**:

Shape

Description automatically generated

Chart, line chart

Description automatically generated

Graphical user interface

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

**6. Contributions**

* Anomaly Detection is ongoing research in Computer Vision (Deep Learning) world.
* Practical implementation of the techniques we have learned in the class like convolution network, drop outs, fully connected networks, pooling, regularizers and etc
* Apart from the convolution networks, here I tried implementing the I3D (Inflated 3 Dimensional networks). This is currently being used in the real world for any type of video classification models. The advantageous thing about I3D is that It uses 3D convolution to learn spatiotemporal information directly from videos

**7. References**

[1] Reynold Cheng Dmitri V. Kalashnikov Sunil Prabhakar Department of Computer Science, Purdue University

[2] [YuTian](https://paperswithcode.com/author/yu-tian), [Guansong Pang](https://paperswithcode.com/author/guansong-pang), [Yuanhong Chen](https://paperswithcode.com/author/yuanhong-chen), [Rajvinder Singh](https://paperswithcode.com/author/rajvinder-singh), [Johan W. Verjans](https://paperswithcode.com/author/johan-w-verjans), [Gustavo Carneiro](https://paperswithcode.com/author/gustavo-carneiro) – ICCV 2021

[3] R.Cheng, D. V. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. Technical Report TR 02-026, Department of Computer Science, Purdue University, West Lafayette, IN 47907, Nov. 2002.

[4] R. Cheng, S. Prabhakar, and D. Kalashnikov. Querying imprecise data in moving object environments. In ICDE’03, Proc. of IEEE Int’l Conf. on Data Engineering, Mar 5–8 2003