Real-Time Anomaly Recognition Through CCTV Using Neural Networks

Problem Statment/Summary

This Paper discusses the need for automated anomaly recognition systems for surveillance videos due to the difficulty in manual monitoring. Anomaly Recognition System is defined as a real-time surveillance program designed to automatically detect and account for the signs of offensive or disruptive activities immediately. Deep learning techniques are used to create a better security system, and two different neural networks, CNN and RNN, have been used for this purpose. CNN is used to extract advanced feature maps from the available recordings, and RNN is used to provide the sense to the captured sequence of actions/movements in the recordings. The output of the system is used to perform real-time surveillance on the CCTV cameras of different organizations to avoid and detect any suspicious activity.

- *InceptionV3- a pre-trained model (CNN)*
- *LSTM cell (RNN)*
- *classification of the video into the 13 groups (12 anomalies and 1 normal).*

Methodology/Significance

The Anomaly Recognition System uses a combination of convolutional and recurrent neural networks. The first neural network is a pre-trained convolutional model called inceptionV3, which extracts high-level feature maps from the surveillance videos to reduce input complexity for the second neural network. Transfer learning is utilized to re-train the pre-trained model for the weights of new classes. The second neural network is a recurrent neural net that extracts meaning from the chain of actions in a fixed time duration and classifies the segments of videos as either a threat or safe. This system aims to automate the process of detecting and classifying suspicious activities in real-time surveillance videos with high accuracy.

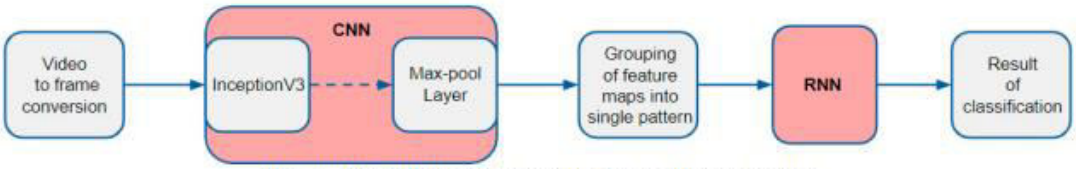


Fig. 1. Workflow of Anomaly Recognition System

Data Set

The UCF-Crime dataset Abuse,Burglar, Explosion, Shooting, Fighting, Shoplifting, Road Accidents, Arson, Robbery, Stealing, Assault, Vandalism, and Normal is a collection of 1800 videos, which includes 950 unaltered real-world surveillance videos containing anomalies and 940 normal scenarios. Text queries were used to scrape these videos from websites like LiveLeak and Youtube with minor alterations for each anomaly. The dataset includes videos in different languages to increase its size. To ensure quality, the dataset was reviewed by 10 trained annotators with varying degrees of expertise in computer vision. Recordings that were manually altered, hoax recordings, news collected, non-CCTV camera captured, or captured by a portable recording camera and containing aggregation were removed. Additionally,

recordings in which the anomaly wasn't clear were also removed.

Table 1. Comparison of different dataset studied with our dataset.

	Video Count	Average Frames	Length
UMN	5	1290	5 min.
UCSD Ped1	70	201	5 min.
UCSD Ped2	28	163	5 min.
Avenue	37	839	30 min.
Subway Entrance	1	121,749	1.5 hours
Subway Exit	1	64,901	1.5 hours
BOSS	12	4052	27 min.
Ours	**1800**	**7247**	**128 hours**

Results

There are 6 models were train with different hyper parameter and data set distribution in which Model 6 has comparatively the best overall performance amongst all the six models that have been implemented during the course of this work as shown in Table that present the results obtained by six models that have been trained so far. Furthermore, they have tested model 6 on the self-collected dataset to examine its application in real time scenarios

Table 2. Performance Comparison of all models.

Models	train_loss	train_acc	val_loss	val_acc	Overfitting
Model 1	0.0652	0.9308	0.0142	0.9630	Maximum
Model 2	0.1819	0.9033	0.0793	0.9896	Considerable amount
Model 3	0.0062	1.0000	0.1333	0.8405	Reduced
Model 4	0.0248	0.8458	0.0569	0.4850	Reduced
Model 5	0.0148	0.9993	0.1341	0.9690	Reduced
Model 6	0.0098	0.9999	0.1548	0.9723	Least

Table 3. Details about the Optimized Model.

	Values
Categories Identified	Abuse, Burglar, Explosion, Shooting, Fighting, Shoplifting, Road Accidents, Arson, Robbery, Stealing, Assault, Vandalism, Normal
Chunk Size	8 frames
Optimizer	Stochastic Gradient
Error Function	Categorical Cross-Entropy
Regularization	Regularizers.l2 (0.01)
Activation Functions	Relu, Sigmoid, Softmax
Augmentation	Horizontal Flip

Link : https://www.sciencedirect.com/science/article/pii/S1877050920315349?ref=pdf_download&fr=RR-2&rr=7b5994e0cdb24d9f

---

Suspicious Activity Detection in Surveillance Footage

Problem Statment/Summary

It is important to develop systems that can detect suspicious activities in surveillance footage to minimize the risk to human life. The traditional method of manual surveillance was tiring and not effective in detecting uncommon suspicious activities. Intelligent surveillance systems have been introduced to tackle this problem. The focus of this approach is to detect potential gun-based crimes and abandoned luggage on frames of surveillance footage, which could lead to high-risk situations. To achieve this, a deep neural network model has been developed that can detect handguns in images and a machine learning and computer vision pipeline that can detect abandoned luggage. These systems can help identify potential gun-based crimes and abandoned luggage situations in surveillance footage, which can prevent harmful events from occurring.

- *Gun based Crime Detection*
- *Abandoned Luggage Detection*

Methodology/Significance

Faster R-CNN is a popular object detection model that uses a region proposal network to generate region proposals and a separate network to classify and refine the proposals. Inception v2 is a convolutional neural network that is used for feature extraction in the Faster R-CNN model. The MS-COCO dataset is a large-scale dataset that contains a diverse set of objects in real-world images, and the pretraining on this dataset allows the detector to learn general object detection features. The combination of Faster R-CNN and Inception v2 is known for achieving good accuracy on object detection tasks while still maintaining relatively fast inference times.

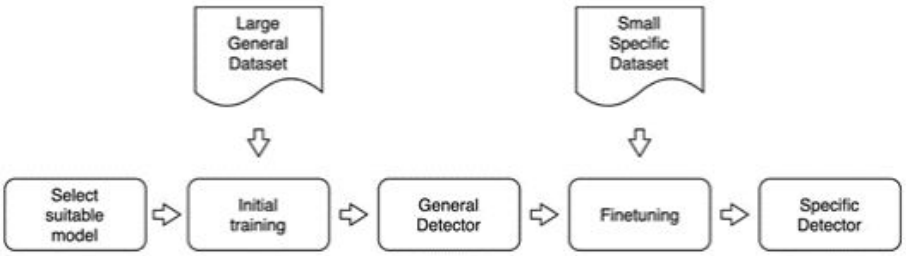


Fig. 1: General workflow of developing a specific object detector.

Data Set

The MS-COCO dataset (Microsoft Common Objects in Context) is a large-scale image recognition dataset that contains over 330,000 images, each of which is labeled with object annotations. The annotations cover a diverse range of object categories, such as people, animals, vehicles, household objects, and more. It is commonly used for training and evaluating object detection, segmentation, and captioning models. The annotations in the MS-COCO dataset are provided in the COCO annotation format, which is a JSON file that contains detailed information about each object in an image, such as its category, bounding box coordinates, and segmentation mask. The dataset is split into three subsets: train, validation, and test. The train set contains over 118,000 images, the validation set contains over 5,000 images, and the test set contains over 40,000 images.

Results

This gun detector's performance was evaluated using the test data we stated before. the models training accuracy was 91.3% and testing accuracy was 89.4% as per the split of data used to train the model. Since the previous mentioned accuracy includes the localization of the object too which differs from the purpose, detecting a gun in a frame of a surveillance footage is sufficient, they focus on the classification part of the object detector rather than the localization of the detected object (bounding box).

	Class 1 - Gun (Predicted)	*Class 0 - Other (Predicted)*
Class 1 - Gun (Actual)	302	2
Class 0 - Other (Actual)	33	271

The model's performance was tested on various PETS 2006 sequences, which included different scenarios, numbers of individuals, and types of baggage. Sample detections performed by the pipeline are shown in Figure 3. Each scenario was captured by multiple cameras with multiple actors involved. The image sequence captured from the camera closest to the abandoned luggage was used for testing purposes. The subjective difficulty of detecting the abandoned luggage in each sequence was defined using stars on the PETS 2006 Benchmark Data website, with one star being the easiest and five stars being the most difficult. Computation time is a crucial factor in video surveillance models. The model's computation time was calculated by dividing the amount of time taken for the entire video's inference

by the number of frames rendered from the dataset.

The method proposed in this study for detecting abandoned luggage does not address issues like identification of objects in sudden changes of illumination therefore, more studies on this line can be carried out for the development of the subject area.

Link: https://sci-hub.se/https://ieeexplore.ieee.org/document/8959600

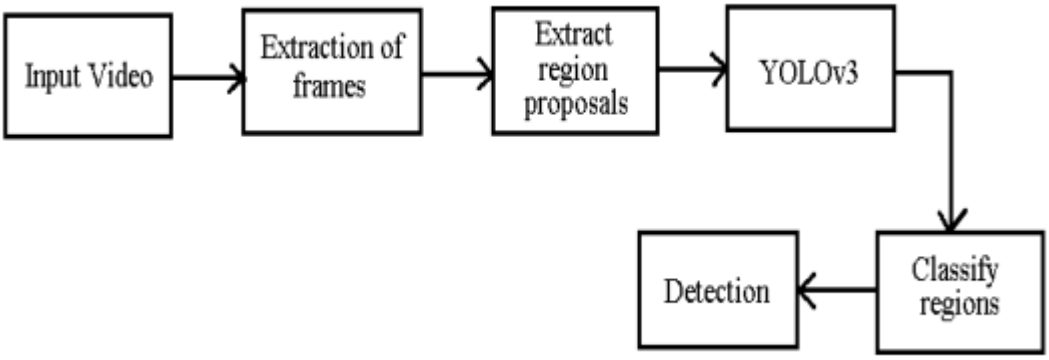
Suspicious Activity Detection from Videos using YOLOv3

Problem Statment/Summary

In the field of Computer Vision and Artificial Intelligence, the detection of human activities in videos is an increasingly important area. A video processing system can automatically analyze video sequences and make intelligent decisions about the actions taking place within them. One particular application of this technology is in the detection of suspicious activities in various environments. To accomplish this, the video is first converted into frames, and then the activities of people within those frames are analyzed. However, the detection of human bodies can be a challenging problem due to the non-rigid nature of human anatomy and the many environmental factors that can affect detection, such as lighting conditions and varying poses. In a recent study, the YOLOv3 algorithm was utilized to detect suspicious activities, such as bag-snatching and lock-breaking, with a high degree of accuracy and processing speed.To locate areas of interest in a video and identify suspicious human activity, detection and tracking methods are often used in tandem. Initially, human detection is performed on each frame of the video, followed by tracking of the identified humans across multiple frames. While current surveillance systems rely heavily on human operators, an algorithmic approach can improve the detection, recognition, and tracking accuracy. To support the development of such an algorithm, a dataset has been prepared. Effective object or human detection, recognition, and tracking can help prevent unwanted events. Thus, the goal is to create a computerized system for automatic human action recognition in suspicious movements that is robust, fast, and accurate across a variety of environments.

Methodology/Significance

The purpose of suspicious activity detection is to identify actions that appear abnormal or suspicious. The system block diagram in Figure 1 outlines the steps involved in this process. First, video data of anomalies performed by different individuals in various backgrounds is collected. This data is then converted to frames at a rate of 30 frames per second and resized as required. Next, the frames are annotated based on regions of interest, followed by training of the dataset using the YOLOv3 model. Finally, the YOLOv3 model is utilized to detect suspicious actions in the input data during testing, distinguishing between normal and abnormal activity.



Data Set

The UMN dataset is a collection of images and videos developed by researchers at the University of Minnesota. It consists of over 4,000 images and 100 video sequences that are captured from a stationary camera in indoor and outdoor environments. The dataset is intended to be used for testing and developing algorithms related to object detection, tracking, and recognition. The images in the dataset are annotated with bounding boxes and labels for the objects in the scene, such as pedestrians, vehicles, and bicycles. The video sequences are also labeled with ground truth data for the motion and location of the objects in the scene over time. The UMN dataset has been used in a wide range of computer vision and machine learning research projects, including object detection, tracking, and recognition, as well as activity recognition and anomaly detection in video surveillance.

Results

Detecting suspicious activity in video data is a complex and challenging task. The framework used for this purpose was executed in Google Colaboratory, which was connected to a Python 3 Google Compute Engine backend with a GPU. The total RAM size used during execution was approximately 0.82GB. The time taken to detect each image ranged from 0.056 seconds to 0.060 seconds, with an average detection time of 0.057 seconds. This indicates that the detection speed of the model is quite fast.

Precision	F1 score	Accuracy
93.10%	96.42%	95%

TABLE II: Comparison with previous study

Model	Precision
BE+OS+FE+AR [14]	85%
Our proposed model	93.10%

The detection of suspicious actions in a security system is a complex task due to the wide variety of human behaviors in natural environments. The proposed system achieved an accuracy of approximately 95%. YOLOv3 outperformed Faster R-CNN in terms of processing time for image detection. However, the current feature extraction method provides accurate results only in a controlled environment. To improve the results, better feature extraction methods can be incorporated.

One limitation of the study was the small amount of data in the training set, which resulted in some mismatch between the test results and the ground truth. To obtain better detection and make the model more practical, future work should focus on extending the training dataset by including suspicious videos of different activities and resolutions. Additionally, more sophisticated algorithms can be designed for

real-time applications.

Link : https://sci-hub.se/https://ieeexplore.ieee.org/document/9342230