JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 3, 2022, p. 2988-2998
https://publishoa.com
ISSN: 1309-3452

Abnormal Event Detection in Video Surveillance Using Yolov3

G.Balamurugan

Assistant Professor
Department of Computer Science and Engineering Manakula Vinayagar Institute of Technology
Pondicherry University, Pondicherry
balamurugancse@mvit.edu.in

Ravi.G

UG Scholar
Department of Computer Science and Engineering Manakula Vinayagar Institute of Technology
Pondicherry University, Pondicherry
ravigovind.pooh@gmail.com

Shanthakumar D. R

UG Scholar
Department of Computer Science and Engineering Manakula Vinayagar Institute of Technology
Pondicherry University, Pondicherry
shanthakumarsk1801@gmail.com

Chandru. J

UG Scholar
Department of Computer Science and Engineering Manakula Vinayagar Institute of Technology
Pondicherry University, Pondicherry
cjayakumar904@gmail.com

Received 2022 April 02**; Revised** 2022 May 20**; Accepted** 2022 June 18**.**

ABSTRACT

The importance of the automated detection of abnormal occurrences within video streams has grown in tandem with the proliferation of video surveillance equipment. When compared to the typical course of events, an abnormal occurrence may be thought of as a departure from the norm. Despite this, the ratio of normal to abnormal occurrences is highly unbalanced due to the fact that abnormal occurrences do not happen very often. A technique for detecting video abnormal events that is based on CNN (convolutional Neural Networks) as well as instance based learning has been suggested. This approach was developed in detection to the need that video abnormal events be localized in pixel-level regions. First, the Gaussian background models are used to precisely pinpoint the moving targets inside the movie, and then, using an image processing approach, the associated areas of the pinpointed moving objects are acquired. Finally, the pre-trained is put to use in order to extract features from linked areas, which are then utilized to generate multiple kernel learning packages. In the end, the multiple instance learning model learns using the normalized set kernel approach, and pixel-level predictions are made. Deep learning and the You Only Look Once v3 (YOLOv3) object detection technique are going to be combined in this model for the detection of highway accidents. According to the findings of the experiments, the technique of video anomaly detection that is based on CNN and sparse representation learning is able to precisely discover abnormal occurrences in the area that is comprised of pixels. Using this as a use case, the purpose of this thesis work was to provide an initial solution for the same problem using deep learning techniques. This was done to avoid the need to involve human resources in the monitoring of any abnormal activities that were spotted in the live broadcasts from the monitoring system.

Keyword: Abnormal Detection, CNN, Deep learning, YOLOv3.

1. Introduction

According to the statistics, upwards of 1.25 million individuals lose their lives as a direct consequence of being involved in a car accident each year. This makes the issue of road accidents a highly significant and high ranking public health matter. Different risk factors, such as speeding, driving under the influence of alcohol or drugs, driving while distracted, driving in an unsafe vehicle, failing to obey traffic laws, and, most critically, failing to provide proper emergency treatment after an accident. Accidents might become more serious if there is even the slightest delay in locating the victim and delivering emergency treatment for them. Because of the progress that has been made in areas such as artificial intelligence, machine learning, and deep learning, we are now able to give our gadgets an increased level of intelligence. Cameras for monitoring traffic have already been set up in the vast majority of the city's neighborhoods. This article was inspired by the concept of using a statistical approach of deep learning to identify any form of collision in a live stream using a convolution neural network. The paper's motivation comes from this concept. The YOLO (You Only Look Once) method, which is part of the Deep Learning Technique, is used in conjunction with Convolutional Neural Networks to perform the thing Detection. Throughout the course of this study, the COCO dataset is used, which has the pre - trained models weight, and its threshold value is around 0.3. Object correctly spotted in both the photos and the videos. An application for Android is developed that can recognize many different kinds of items and provide the user with verbal feedback on its findings. It is said that object detection using the YOLO method is quicker in comparison to other classification algorithms. Additionally, it is stated that although it does make mistakes in localization, it predicts fewer false positives in the background. According to the technique that has been suggested, the typical IoU is 83.19 percent, and the typical mAP is 98.14 percent. Objects detection in stack photos has the potential to address numerous issues that arise in the field of retail sales. These issues include monitoring the quantity of goods that are stocked on shelves, filling in gaps in inventory, and continually matching planograms. In recent years, there has been widespread use of the deep learning technology that is quickly advancing. Deep learning is a relatively new academic discipline that aims to perform data analysis and interpretation by modelling and imitating the manner in which the human brain works. In deep learning, the primary focus is on the process of "training" multilayer CNN architectures by feeding them massive amounts of data in order to establish input-output associations. The specific challenge at hand will influence the number of stages included inside a model, as well as the number of nodes contained within each layer, the manner in which the neurons are linked, and the manner in which the functions are emulated and decided. Large amounts of data are used in the process of updating the weights and biases of each layer.

2. Related Works

Reconstruction-based anomaly detection techniques were put forward by Hasan and others. These techniques employ video sequences or bespoke features (such as low-level trajectory information) as the input and extract high-level feature maps to model normalcy. To understand the temporal regularity of the typical occurrences, the reconstruction error may be minimised. The abnormal patterns will result in a larger reconstruction error since these models can only learn the structures that are present in the typical training data set. As a consequence, the accuracy of the reconstructions may be used to distinguish between abnormal events.

According to W. Liu et al., the computational time of abnormal events is not always more than that of the normal occurrences because of the enormous capacity of deep neural networks. As a consequence, [2] suggested an anomaly detection technique based on predictions. This technique uses a U-Net architecture to predict the next image from the previous consecutive frames and then compares the forecast with subsequent frames to spot abnormal occurrences.

The methodologies proposed by Mahadevan et al. primarily concentrate on acquiring optical flow and the temporal consistency, but they neglect to take into account an essential component known as the appearing abnormality cue, which is equally significant. Because of this, they are unable to detect certain anomalous things, which are distinguishable from regular objects in appearance but do not entail motion outliers. These objects have a typical appearance but do not involve any motion outliers.

The authors were inspired by the promise of sparse coding-based suspicious detection, according to the study [4]. As a consequence, they devised a Temporally-coherent Sparse Coding (TSC), wherein the researchers used the same reconstruction coefficients to build similar surrounding frames for encoding. Then, using a special arrangement of stacked recurrent neural networks, we mapped the TSC (sRNN). Some of the developments made by the work include the ones listed below:it proposes a TSC that is capable of being recorded to a sRNN, which makes it easier to optimise the parameters and speeds up the uncertain prediction. ii) Create a very big datasets that is even greater than the total of all the other datasets that are currently available in order to search for unusual behavior.

A method that is useful for detecting abnormalities in movies was detailed in a research article published by Springer [5]. The promise of convolutional layers for image processing applications has been demonstrated by recent approaches to convolutional neural networks.particularly in the context of photographs. Convolutional neural networks, on the other hand, are supervised, meaning they need labels as a kind of learning signal. A spatiotemporal architecture for the detection of suspicious activity in movies including crowded settings has been presented by the authors as well as others.

The paper proposed the end-to-end easily trained complex Convolutional Long Short-Term Memory (Conv-LSTM) networks. [6]. [Citation needed] These Conv-LSTM networks are able to anticipate the progression of a video sequence based on a very small number of the frames that were used as input. Consistency ratings are produced using a set of estimations' reconstruction errors, with irregular video sequences producing lower consistency scores as they become more and more disconnected from the genuine sequence over time. The models have made use of a structural framework to explore the special effects reinforcement has on the stages of studying representations that carry greater weight.

According to [7], the solution to this challenge may be found by first developing a generative model for persistent motion patterns. This model should make use of many resources and should have very little supervision. In particular, the study presents two approaches that are based on auto encoders because of their capability to function with a little amount of or even no supervision at all. The first strategy is to investigate the fully linked auto encoder after first using the usual handmade spatio-temporal local features. Building an end-to-end learning structure composed of a convolutional feed-forward auto encoder is the second phase. As a result, the classifiers and local features may be learnt simultaneously.The model that has been suggested is capable of capturing the regularities that are present in a variety of datasets.

3. Proposed Methodology

In the system that we have developed, the CNN, also known as the convolution neural network, has been used for the purpose of identifying unusual patterns of behaviour. It is very necessary to identify the temporal data shown in the video in order to accurately classify the unusual actions. In recent years, the most common use for CNN has been the extraction of essential information from each frame of a video. CNN is the only algorithm that even comes close to being adequate for this task. CNN has to be able to recognize and extract the required characteristics from the screen of videos in order for the classification of the provided input to be effective. Because of this, CNN needs to be able to know which features to extract from the frame of videos. In order to recognize and categories the items, the suggested system was built using YOLOv3 as the underlying technology. Data gathering, data modelling, training and testing of the model, and performance analysis are the processes that are involved in the system that is being suggested.

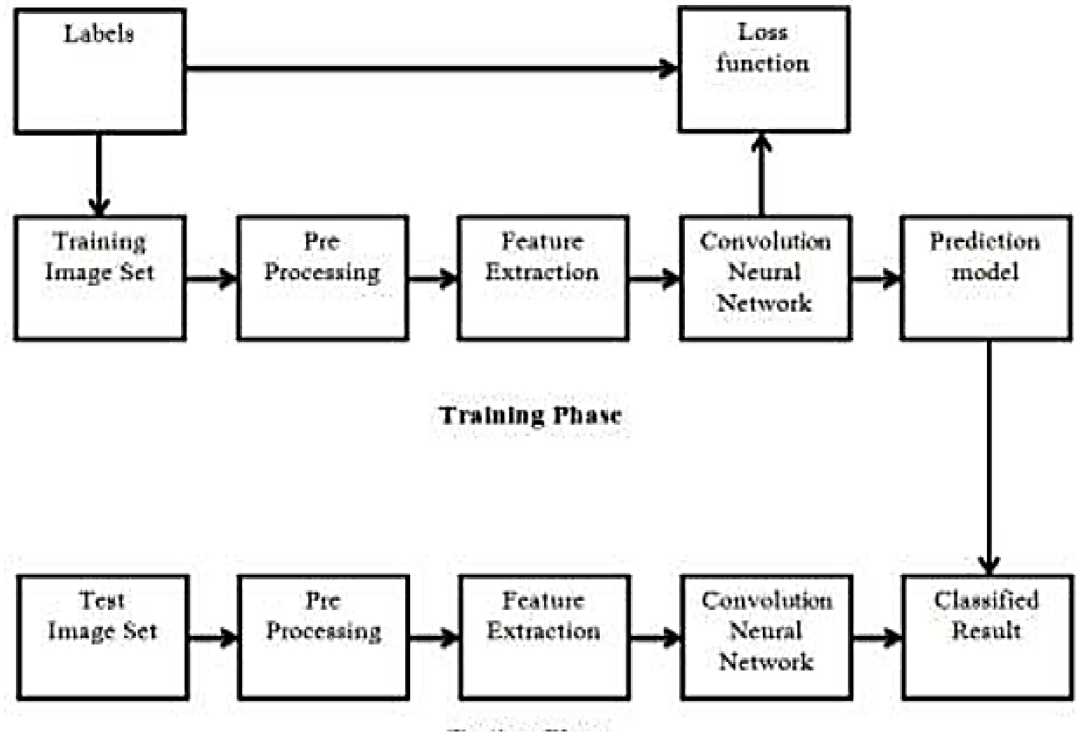


Figure 1. Proposed Architecture

Data modeling

Object detection and categorization are both accomplished with the help of YOLOv3. Darknet-53 consists of 106 layers, including 53 convolutional layers on top of 53 more layers, for a total of 106 layers.

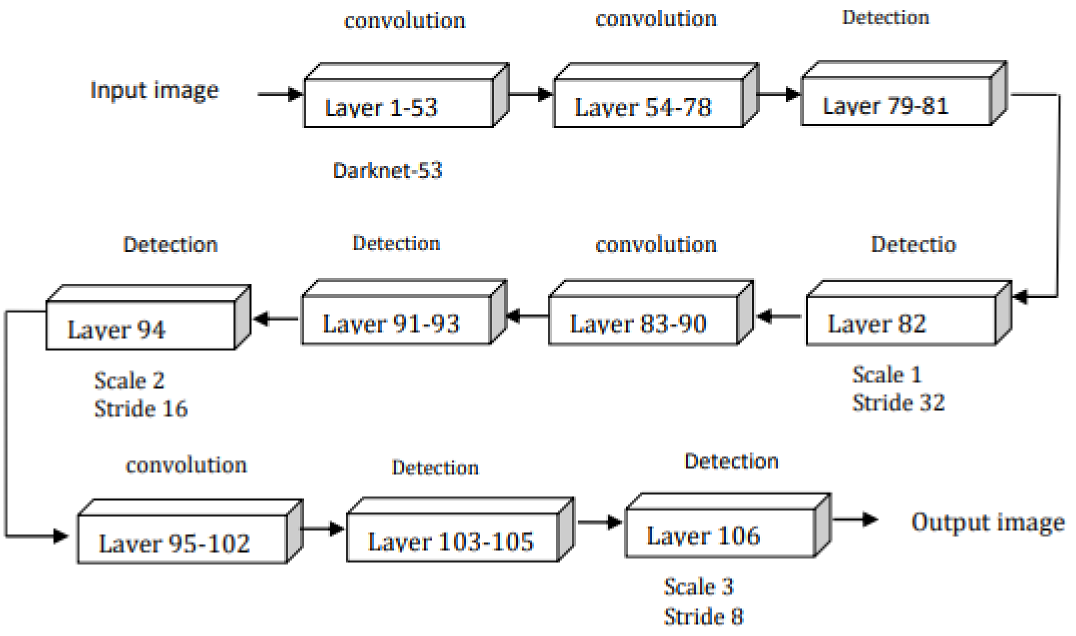


Figure2: YOLOv3 Architecture

The picture that is being read in is a 416 by 416 RGB image. Each convolution layer in Darknet-53 is then followed by a batch normalizing layer and a LeakyReLU layer. YOLOv3 performs detections at three distinct scales simultaneously. At layers 82, 94, and 106, respectively, the network performs a stride 32, 16, and 8 down sampling on the input picture. The final feature maps have dimensions of 13 by 13, 26 by 26, and 52 by 52. After that, each detection is carried out using a 1x1 detection kernel, which produces the detection feature maps 13x13x255, 26x26x255, and 52x52x255 respectively. In this case, the responsibility for identifying big things falls on the 13x13 grid, while the 26x26 grid is responsible for detecting medium-sized objects, and the 52x52 grid is responsible for detecting little objects. A picture is predicted to have 8112 bounding boxes at size (52x52), 507 bounding boxes at scale 1 (13x13), and 2028 bounding boxes at scale 2 (26x26) using YOLOv3. These are screened utilizing the two approaches that are following.

Proposed Algorithm: Convolution Neural Network (CNN)

Step 1: a picture or video will be provided as the input.

Step 2: a variety of filters are applied to the data that was provided in order to produce a feature map.

Step 3: Afterwards when, a ReLU function is used to increase the output's nonlinearity.

Step 4: Every one of the convolution layers receives a pooling layer.

Step 5: the technique involves the compression of all of the combined pictures into a single lengthy vector.

Step 6: The vector is then sent into the algorithm so that a fully connected artificial neural network may utilise it in the next stage.

Step 7: Processes the features across the network. When all is said and done, the "vote" of the classes is delivered via the completely linked layer.

Step 8: The last step of the training process involves training via both forward and reverse propagation over a number of epochs. Repeating this process until we have a well-defined neural network with training weights and feature detectors is what we do until then.

Steps for Videos to frame conversion

An effective motion auto encoder that outputs the RGB difference and accepts successive frames as input

Step 1: An picture or video is provided as input.

Step 2: After applying several filters to the input, a feature map is produced.

Step 3: To further boost non-linearity, a ReLU (rectified linear)function is next used.

Step 4: After that, all and every convolution layer is given a pooling layer.

Step 5: The method merges all of the pooled pictures into a single long vecto

Step 6: The last step involves feeding the algorithm's vector into a fully linked artificial neural network.

Step 7: Utilizes the network to process the features. The "vote" of the courses is delivered via a fully linked layer at the conclusion.

Step 8: This last step involves training across several epochs using both forward and reverse propagation. This cycle is repeated until the neural network is well-defined, with training weights & feature detectors.

MODULE DESCRIPTION

- Data Collection
- Preprocessing
- Noise removal
- Resizing
- Binary conversion
- Segmentation
- Feature extraction
- Classification

Data Collection:

Data is first retrieved for a variety of websites and apps for social media platforms based on the factors that differentiate each of these types of platforms.

Preprocessing:

After that, in order to clean up our dataset, we will go through a series of pre-processing procedures such as removing noise, resizing it, converting it to binary, and scaling it grayscale.

Noise removal:

The video that was submitted has had its noise reduced. Filtering is the most important part of the denoising process in image processing. Filters such as the average filter, the median filter, the Wiener filter, and the Kalman filter are often applied in the process of noise reduction.

Resizing:

Picture remapping is a possibility when correcting for vision sensor or flipping an image, but image scaling is necessary when we really need to increase or decrease the number of pixels. When we are doing lens distortion corrections, image remapping could be used.

Binary conversion:

Since the only color combinations that can still be represented precisely twice each are black and white, binary images are made up of pixels which can only be either of these two colours. Binary pictures are also known to as bi-level or image pairs on occasion. This implies that every single pixel is stored in the system as a single bit, or alternatively, as a value that may either be 0 or 1.

Segmentation:

Image segmentation is a fundamental technique that involves isolating a digital image into various segments. These segments may be anything from few pixels to whole scenes (sets of pixels, also recognised as image objects).

Feature extraction:

The process of separating and condensing an initial collection of raw data into more easily manageable groupings is known as feature extraction. This process is a component of the dimensionality reduction approach.

Classification:

Classification is the process of organizing and naming distinct groupings of pixels or vectors included within a picture in accordance with predetermined criteria and guidelines.

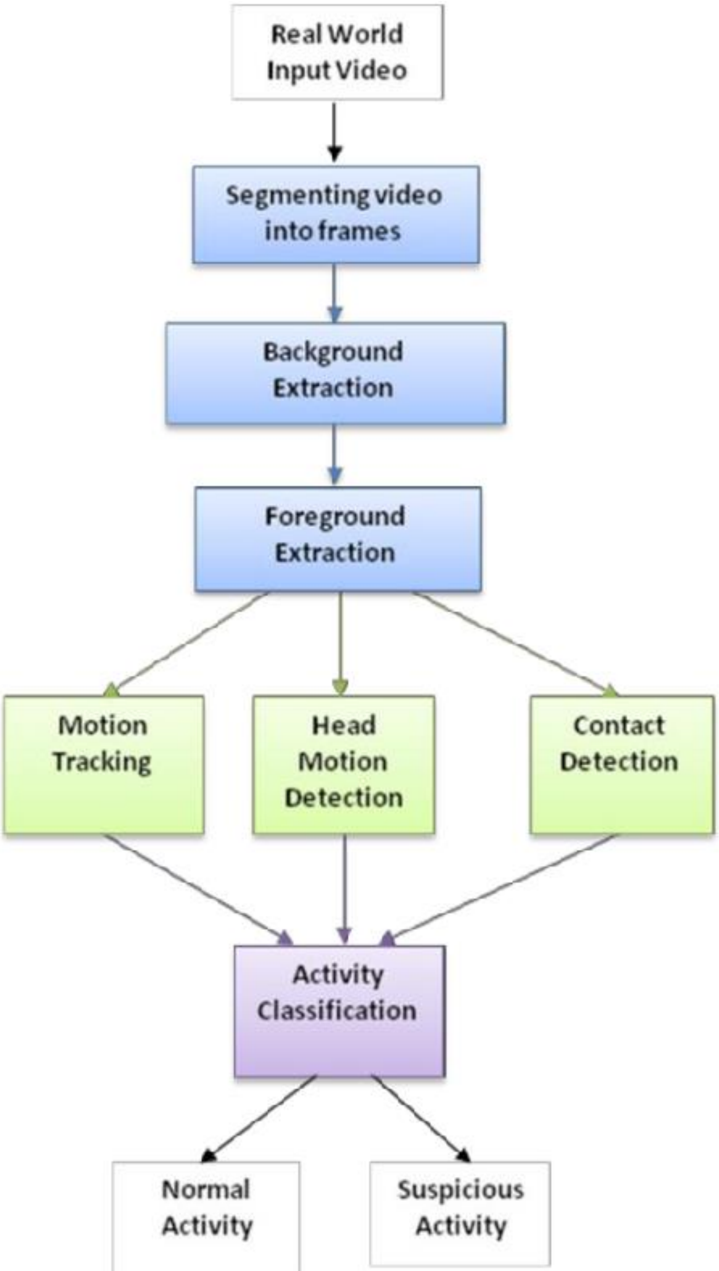


Figure 3. Data Flow Diagram for Abnormal Detection

4. Experimental Results

Dataset Description

The efficiency of the suggested method has been validated by the use of the two widely used datasets. The initial set of datasets are integrated using datasets such as SumMe, MED, and TVsum. In this particular piece of research, we refer to this as the combined dataset. The datasets are quite comparable, whether it in terms of the visual styles or the contents. This method covers the problem of insufficient data for training, as well. In a nutshell, the merged dataset consists of 235 films, each of which has an average runtime of two minutes and a frame count of around three thousand. In this step, the combined dataset is partitioned into two sets: the testing set, which contains 55 films, and the training set, which has 180 videos.

Training and testing

The training was carried out with the assistance of Google Collaborator, which offered a Tesla P4 GPU to facilitate the quicker and more effective training of the network. The batch of photos in the training dataset totals 1500 in total. The dataset was trained for a total of 10000 iterations, which was calculated by multiplying the number of total classes by 2000. The entire amount of time that was necessary to train the network using the setups described above was around 14-16 hours. After 10000 iterations, the weights that were created this way were used in order to identify and examine the performance. During the course of testing, a batch of one hundred photographs was evaluated. It has a higher degree of accuracy when detecting class labels and bounding boxes.

Performance analysis

Accuracy, recall, mean absolute precision, index of uniformity, and f1 score are the test criteria. The determination of these performance indicators is accomplished by using true positive, false positive, and false negative numbers. Precision may be defined as the proportion of correct positive class predictions relative to the total number of correct positive class predictions.

$$\text{Precision} = TP / (TP+FP)$$

The number of correct predictions of class made out of the total number of results forecasted is referred to as the recall.

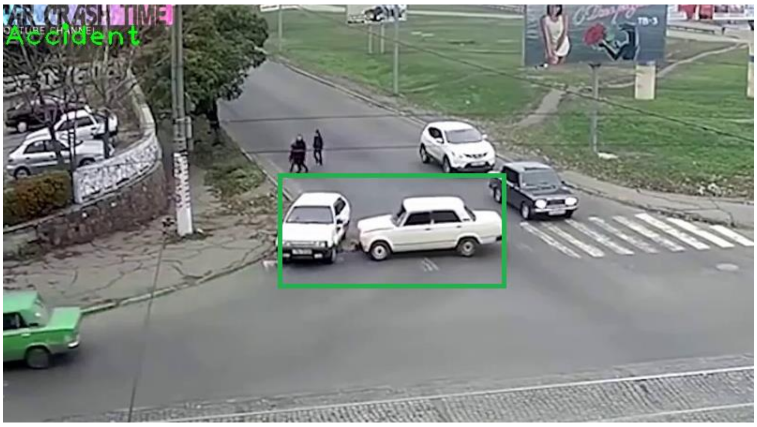
$$\text{Recall} = TP / (TP+FN)$$

The F1 score is a weighted average that takes into account both accuracy and recall.

$$\text{F1-score} = 2 * (\text{precision}*\text{recall}) / (\text{precision} + \text{recall})$$

The Confidence/Model scores are obtained by using a real-time feed to check for crash avoidance utilizing models and tensor flow. If the score hits a threshold of 0.7 or more, then an accident has been identified, and an SMS message has been issued. The mAP value of the optimized yolo is another significant signal that should be considered. The mean of the class's average precision is denoted by the term "mean average precision." Because there is only one class that has to be identified in this study, the mean average precision is the same as the average precision (AP).

The efficiency of the suggested method has been validated by the use of the two widely used datasets. The initial set of datasets are integrated using datasets such as SumMe, MED, and TVsum from www.kaggle.com. The datasets are quite comparable, whether it in terms of the visual styles or the contents. This method covers the problem of insufficient data for training, as well. In a nutshell, the merged dataset consists of 235 films, each of which has an average runtime of two minutes and a frame count of around three thousand. In this example, the combined dataset is partitioned into a testing set consisting of 55 movies and a training set consisting of 180 films, as can be seen in figures 4 and 5.



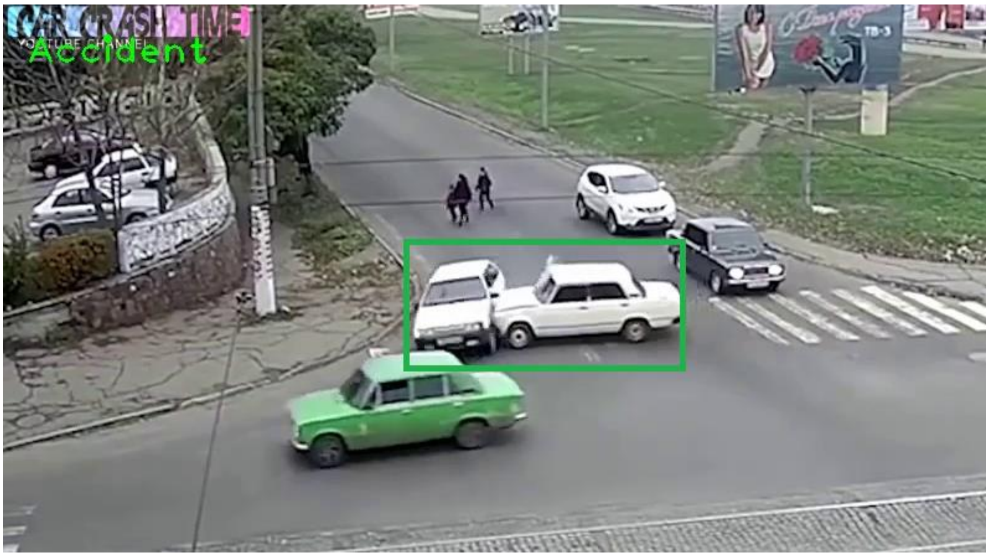


Figure 4. Accident Detection in frame

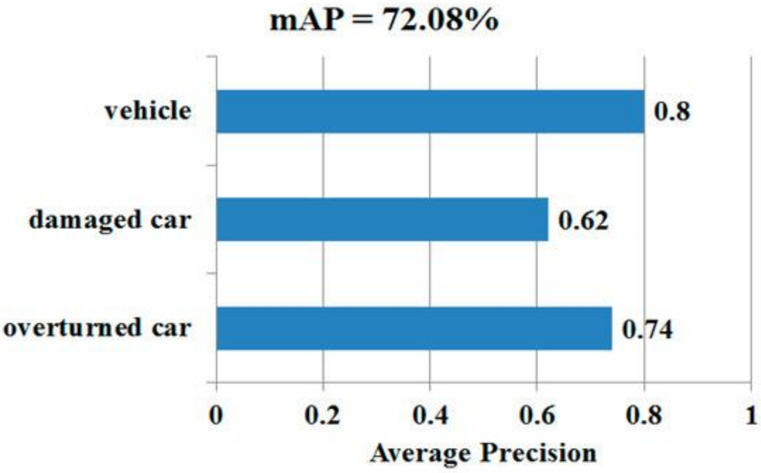


Figure 5: Average Precision

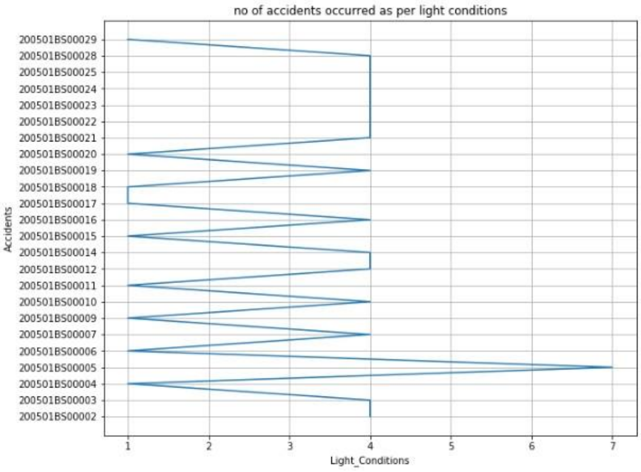


Figure 6: No Of Accidents Occurred As Per Light Conditions

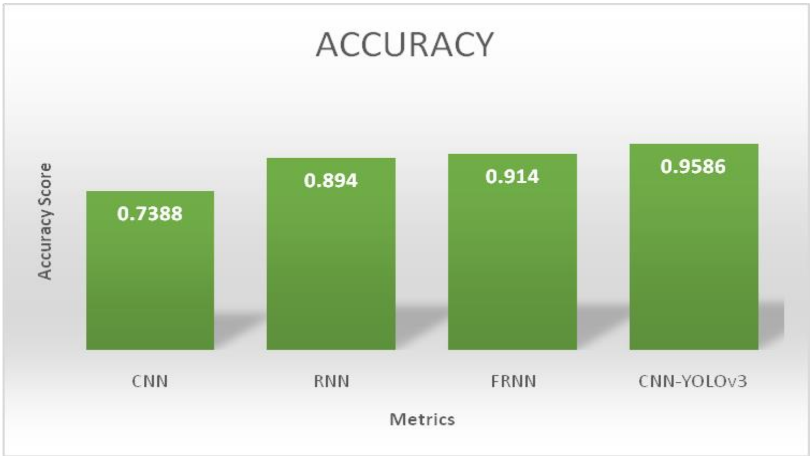


Figure 7: Comparison Graph

5. Conclusion

Combining the YOLOv3 image detection technique with our deep learning model, which we developed, we were able to develop an effective model for the detection and categorization of highway accidents. This system is also not able to identify accidents, but it can also conduct an easy categorization of accidents and make an objective preliminary judgement of the severity of an accident. In addition, this model is able to execute all of these tasks simultaneously. In the meanwhile, the ratio of different classes existing in the whole database may be changed using the YOLOv3 detection technique. By varying the proportion of the various classes, this contributes to the objective of developing a more appealing product. The YOLOv3 detection system can also recognize the edges of a car, which greatly improves classification accuracy. This was made feasible by YOLOv3's capability to elicit a car's boundaries. Additionally, this model was created using photographs taken first from range of angles of dash cams, which is the most crucial starting point in the whole environment. This model was created based on this setting's most important starting point.According to the findings of the experiments, the technique of video anomaly detection that is based on CNN and instance based learning is able to precisely discover abnormal occurrences in the area that is comprised of pixels. Using this as a use case, the purpose of this thesis work was to provide an initial solution for the same problem using deep learning techniques. This was done to avoid the need to involve human resources in the monitoring of any abnormal activities that were spotted in the video stream from the surveillance system.

REFERENCE

[1] M. Hasan, J. Choi, J. Neumann, A.K. Roy-Chowdhury, L.S. Davis, Learning temporal regularity in video sequences, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 733–742.

[2] W. Liu, W. Luo, D. Lian, S. Gao, Future frame prediction for anomaly detection–a new baseline, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6536–6545. [3] V. Mahadevan, W. Li, V. Bhalodia, N. Vasconcelos, Anomaly detection in crowded scenes, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 1975–1981.

[4] W. Luo, W. Liu, D. Lian, J. Tang, L. Duan, X. Peng, S. Gao, Video anomaly detection with sparse coding inspired deep neural networks, IEEE Trans Pattern Anal Mach Intell (2019), doi:10.1109/TPAMI.2019.2944377. 1–1

[5] N. Srivastava, E. Mansimov, R. Salakhudinov, Unsupervised learning of video representations using lstms, in: International conference on machine learning, 2015, pp. 843–852.

[6] F. Tung, J.S. Zelek, D.A. Clausi, Goal-based trajectory analysis for unusual behaviour detection in intelligent surveillance, Image Vis Comput 29 (4) (2011) 230–240.

[7] D. Xu, Y. Yan, E. Ricci, N. Sebe, Detecting anomalous events in videos by learning deep representations of appearance and motion, Comput. Vision Image Understanding 156 (2017) 117–127.

[8] A. Markovitz, G. Sharir, I. Friedman, L. Zelnik-Manor, S. Avidan, Graph embedded pose clustering for anomaly detection, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 10536–10544, doi:10.1109/CVPR42600.2020.01055.

[9] Y. Zhang, H. Lu, L. Zhang, X. Ruan, Combining motion and appearance cues for anomaly detection, Pattern Recognit 51 (2016) 443–452, doi:10.1016/j.patcog. 2015.09.005.

[10] T.-N. Nguyen, J. Meunier, Anomaly detection in video sequence with appearance-motion correspondence, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1273–1283.

[11] G.Balamurugan, J.Jayabharathy (2022). A Comparative Analysis of Event Detection and Video Summarization. In: Hu, YC., Tiwari, S., Trivedi, M.C., Mishra, K.K. (eds) Ambient Communications and Computer Systems. Lecture Notes in Networks and Systems, vol 356. Springer, Singapore. https://doi.org/10.1007/978-981-16-7952-0_54

[12] G. Balamurugan and J. Jayabharathy, “Abnormal Event Detection using Additive Summarization Model for Intelligent Transportation Systems” International Journal of Advanced Computer Science and Applications(IJACSA), 13(5), 2022. http://dx.doi.org/10.14569/IJACSA.2022.0130586.

[13] J.Jayabharathy, G.Balamurugan, R.Vishnu Priya (2021) Abnormal event summarization in video surveillance using hierarchical recurrent neural network, DE, pp 3568–3579.

[14] G, Balamurugan, J, Jayabharathy, An Efficient CNN and BI-LSTM Model for Abnormal Event Detection in Video Surveillance (May 22, 2021). Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021), http://dx.doi.org/10.2139/ssrn.3851212.