“METHODOLOGIES OF VIDEO ANOMALY DETECTION”

A Dissertation

Submitted In Partial Fulfillment Of The Requirements
For The Award Of Degree Of

MASTER OF TECHNOLOGY IN
COMPUTER SCIENCE & ENGINEERING

Submitted by:

RAJIV KUMAR (2K20/CSE/18)

Under the supervision of
Dr. R K Yadav
(Assistant Professor)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering) Bawana Road, Delhi-110042

MAY, 2022

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CANDIDATE'S DECLARATION

I, Rajiv Kumar, Roll No. 2K20/CSE/18 student of M. Tech (Computer Science and Engineering), hereby confirm that the project Dissertation titled "Methodologies of Video Anomaly Detection" that is presented by me to the Department of Computer Science & Engineering, Delhi Technological University, Delhi in substantial fulfilment of the requirement for the granting of the degree of Master of Technology, is genuine and not reproduced from any primary source without appropriate credit. This work has not previously been the basis for the awarding of any Degree, Diploma Associateship, Fellowship or other analogous title or distinction.

Place: Delhi **Rajiv Kumar**

Date: **2K20/CSE/18**

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Bawana Road, Delhi - 110042

CERTIFICATE

I do hereby officially confirm that Project Dissertation labelled "Methodologies of Video Anomaly Detection" that is presented by Rajiv Kumar, 2K20/CSE/18 Department of Computer Science & Engineering, Delhi Technological University, Delhi in substantial fulfilment of the requirement for the award of the degree of Master of Technology, is a documentation of the project work undertaken by the students under my guidance. To the best of my knowledge, this thesis has not been presented in parts or as a whole for any Degree or Qualification to this Institution or elsewhere.

Place: Delhi

Date:

Dr. R K Yadav

Assistant Professor

Department of CSE

ACKNOWLEDGMENT

The success of this initiative involves the support and contribution of several individuals and the organization. I am thankful to everyone who assisted in creating this dissertation.

I convey my deepest thanks to Dr. R K Yadav, my project guide, for giving me the chance to pursue this research under his leadership. His continual words of support have made me aware that it is the method of gaining knowledge which counts more than the ultimate outcome. I am definitely grateful to the panel faculties during all the advancement reviews for their guidance, continual monitoring and for inspiring me to finish my job. They assisted me throughout with fresh ideas, gave information essential and pushed me to finish the assignment.

I additionally thank all my friends and my family for their continuous encouragement.

RAJIV KUMAR

2K18/CSE/18

ABSTRACT

All cities are getting smart with the intervention of latest technologies, their infrastructure is getting upgraded with each day. Critical information is provided to us by these infrastructures. There is growing prevalence of AI in today's world, with the help of which a real-time system can be developed that can assist in detecting crimes as they occur. The surveillance platform's information may include both aberrant and conventional footage. We propose developing an aberrant event identification system based on weakly annotated training videos, and so when such behavior is discovered, suitable action may be taken. For extraction of features, we deployed I3D-Resnet-50, a deep residual model. The Kinetics video action dataset was used to train this network. There are 13 unique abnormalities in our dataset. Crime, Attack, Firing, Burglaries, Thieving, Prison, Fight, Thefts, Breaking and entering, Bomb, Criminal damage, Torture, and Traffic Accident are all unusual incidents. The proposed approach for visual anomaly detection achieves considerable improvements in terms of correctness and recall.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

1. CNN: Convolutional Neural Networks
2. RNN: Recurrent Neural Network
3. C3D: 3D Convolution Network
4. LSTM: Long Short Term Memory
5. AI: Artificial Intelligence

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

Video anomaly detection is described as spotting strange or abnormal occurrences in the videos. It primarily focuses on whether the current frame of the video is unusual or not. There is no correct definition of anomaly but it is viewed as some tragedies or calamity which generally departs from what is typical, customary or anticipated. Detecting anomalous behavior plays a very vital part in intelligent town administration for example criminal probe and transportation monitoring etc.

There has been a spike in the demand for civilian safety which necessitates installation of surveillance cameras in public to follow human behavior and prevent unusual circumstances like bulgari, traffic accidents and criminal activities. Usually there is a requirement of human involvement to monitor the unexpected actions but it is a challenging and time demanding operation.

Hence there is a need to build projects and undertake research in automated video anomaly detection as it would not only lower the human involvement but also enhance the efficiency of the infrastructure to identify crime in a short period of time so that suitable action could be done to avert it. Conventional techniques strive to leverage the trajectory based methodologies.

The primary idea of these approaches hinges on whether the item is following the typical or customary route or not. If it is not follows the regular course then such activity might be classified an oddity. But such approaches demands a substantial amount of labour is required in order to see the things that is of interest, which sometimes may not be suitable for the recording purpose. Not only this, but these tactics shown to be useless when they are applied to new territory since they cannot discover new abnormalities that haven't happened previously.

In some of the last years, all the methods that used deep learning methods has gained so much popularity in the area of computer vision. There are many studies that have been done that uses deep learning methods for solving problem of anomalous behavior detection. Figure 1.1 demonstrate that as the input data size of the dataset increases then the methods that uses deep learning approaches tends to show better accuracy than the conventional based methods.

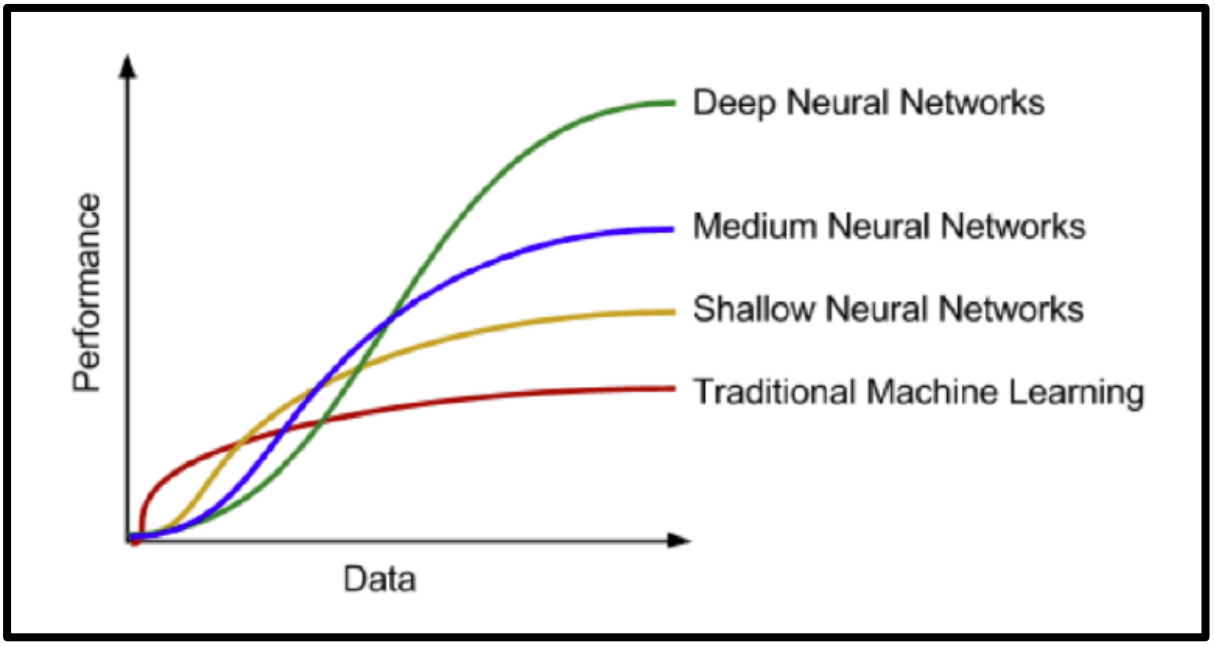


Fig 1.1 Deep Learning against Classical Algorithms.

The problem of video anomaly detection is concerned with the processing of video and therefore there are just two kinds of features that can be extracted by models that uses deep learning methods. First one is spatial features and the other one is temporal feature. There are various deep learning models but CNN or Convolution Neural Networks have been proved to show the most optimal accuracy in spatial data features. In the previous 10 years, numerous methods have been developed that have their significant merits and downside.For example, RNN or Recurrent Neural Network gathers temporal information by addition of transmission mechanism in the hidden layers. However this method is still inefficient when the videos are lengthy. Hence, there is a need to develop model that overcomes the inadequacies of the past techniques.

1.2 CHALLENGES

In contrast to action recognition where events are clearly specified, the characterization of anomalies in video might contain some degree of uncertainty. Anomalies frequently encompass a broad spectrum of activities. In addition, the definition of an anomaly varies in various applications and datasets. In certain circumstances, different authors may identify distinct anomalies as ground truth on a single dataset.

The availability of labeled data, which are utilized to train or to evaluate models for anomaly detection, is frequently a big difficulty. In practice, it may be straightforward to offer sufficiently numerous samples of regular activities, although it is impossible to present all potential instances and forms of deviant activity that might happen in the scenario. As a consequence, it is difficult to train a model in a supervised way to differentiate an aberrant

class from a normal class.

An anomaly detection system is meant to offer real-time and automated alerts when an abnormality emerges in the scene. Furthermore, computational burden and time complexity ought to be handled.

1.3 COMPUTER VISION

Computer vision [2] is the subset of artificial intelligence that lets the computer find knowledge and extract useful information from digital photographs and movies. We can acquire essential information from photos and movies with the assistance of computer vision. Based on the obtained knowledge we may make judgments. In today's world, many public locations have installed CCTV for surveillance purposes, which include airplanes, roads, public transit, shopping malls, banks, and many organizations.

Anomalous behaviors are not only unusual but also difficult to identify and time expensive. CCTV real-time data may be utilized to identify illegal behavior. There are several systems that require human interaction to keep an eye on surveillance recordings. Not all firms are enormous, others have fewer security employees.

It gets difficult for any person to retain attention on more than 20 screens at the same time. So there is a need to design creative algorithms to identify illicit behavior based on patterns [3]. Figure 1.2 depicts the present flow diagram of the monitoring system.

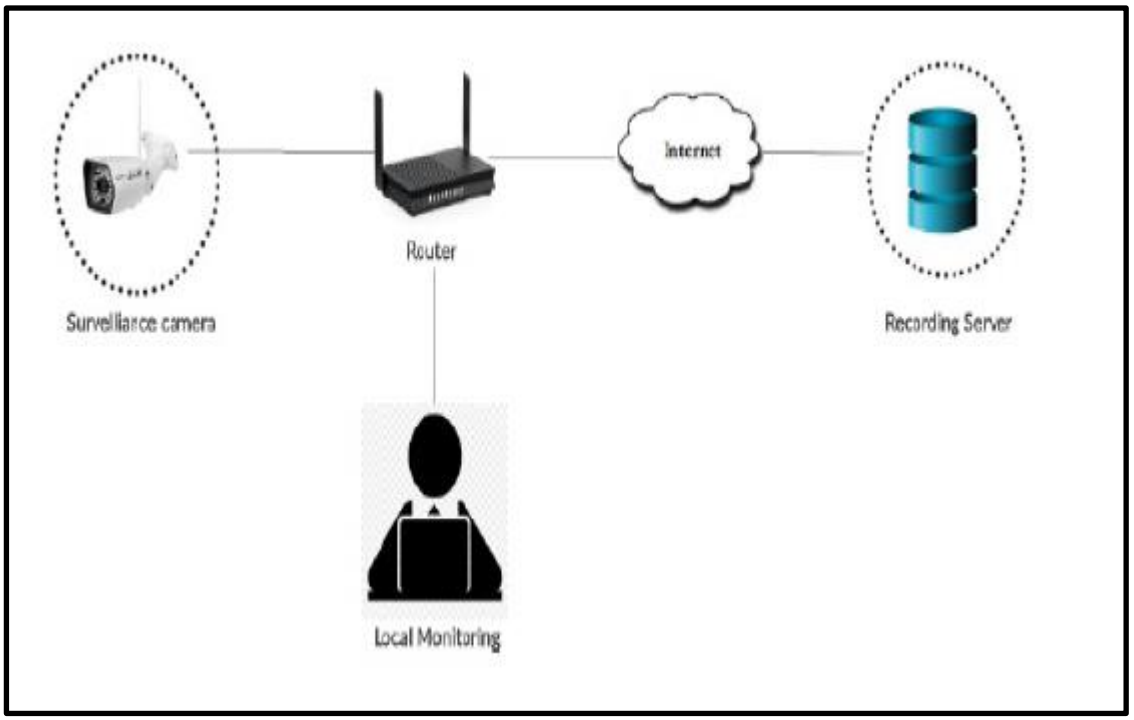


Fig 1.2 Process flow diagram of existing system

We need to build clever programs which will be able to study from trends and recognise illicit actions. Implementing such sophisticated algorithms would drastically minimize human interference in surveillance supervision. Whenever there is abnormality, the system is supposed to label it utilizing the categorization approaches. In practice, any activity that is viewed as normal in one set of circumstances may be labeled aberrant in another set of circumstances. Most of the methodologies presented in the literature uncover irregularity by analyzing unique circumstances, such as those found seldom in films [4, 5, 6].

1.4 Deep Learning

Deep learning is a fast expanding machine learning and AI subject that focuses on applying neural networks to address complex practical difficulties. Deep neural networks provide a broad variety of applications, including visual processing, image processing, and

classification, as well as several modules of self-driving autos. It may be additionally used for video analysis. Deep learning is a fast expanding machine learning and AI subject that focuses on applying neural networks to address complex practical difficulties.

Deep learning is employed in all the video processing research. Due to its high performance it has been extensively used in the area of Artificial Intelligence.

This research mainly focusses on how deep learning has proved to be an efficient approach for solving problems of anomalous behavior detection. Not only this but it also analyses the most commonly and famous dataset that are used in this domain. There evaluation procedure is also discussed.

And after that, a debate incorporating all of the preceding strategies is provided to offer direction and proposed areas for further research.

1.5 Convolution Neural Network

Convolutional Neural Networks [7] is frequently used for feature extraction from pictures but due to lack of dynamic modeling it is surely not suited for video analysis. High Level variation can be replicated using Long short-term memory (LSTM) [8], but minimum motion cannot. In addition, owing to their intricacy, LSTMs are exceedingly expensive to train. The 3D Conv-Net (C3D) [9] is perfect for learning spatiotemporal features. C3D can show temporal data thanks to its three-dimensional convolution and pooling operations, making it better than 2D Conv-Net. The fundamental reason for this is because 2D Conv-Net merely execute spatial pooling and convolution, whereas 3D Conv-Net do these operations both geographically and temporally.

3DConvNet, on the other hand, has one additional kernel dimension. The number of

parameters rises as a consequence, making model training more complicated. There is a requirement to come up with an algorithm that is economical and less complex in training.

CHAPTER 2

PRIOR WORK

2.1 Description Of Some Publicly Available Datasets.

In this Section we will analyze some of the prominent datasets that have been generated for video anomaly detection. It takes a lot of effort to generate the datasets and the datasets we are going to analyze are the product of hard work and all the experiments done by the researchers to answer the real concerts in this area. The surveillance cameras are positioned in a range of areas with varied geographies which assists in creating a broad diversity of the anomaly datasets. A broad variety and quantity of the anomaly footage are provided to the researchers and general public for the study purpose.

The datasets that are available for research, some of them lack in information and are not readily accessible for the study purpose. Hence, a major portion of computer vision researchers are in darkness.

This is due of the fact that the dataset that is created is posted in several of the excellent conferences and is freely viewable on their webpage. The researchers work on this dataset to compare their findings to some of the prior outcomes that was disclosed outside. Hence substantial amount of research can be done in this field by employing these datasets in present application that were unavailable in others. As a consequence of this, there is a

requirement to interpret various abnormal video datasets. The dataset that has been described below provide a broad variety of anomalous dataset that are accessible to the research community. Fig.1.3 illustrates many example frames from several aberrant video datasets.

Fig. 1.3 Anomalies in datasets

UCSD: The UCSD abnormality detection dataset [10] is made using security film captured by the camcorder deployed in two different scenarios of a crowded walking path. It comprises both conventional and exceptional behavior on the pathways, including such wandering pedestrians as well as the activities of riders, skating, cyclists, small buggies, persons in wheelchairs, and so on.

Shanghai Tech Dataset: This is one of the most renowned dataset for video anomaly identification that contain around approximately 13 themes that are not only distinct but also with numerous lighting and camera perspectives. It has around approximately 260, 000

training frames and roughly 128 aberrant behaviors. This dataset offers the pixel level ground truth for the aberrant films featured in this collection. This dataset is deemed ideal if the model is to be deployed in diversity of viewing angle and distinct lightning.

This dataset contains the anomalies that are generated by the jerky motion which includes fist fights and human chases which is hard to find in other datasets. Because of all these small properties, this dataset is considered appropriate for utilization in practical purposes.

UCF Crime Dataset: This Dataset comprises near about 130 hours of videos that contains anomalous behavior. It also consists of uninterrupted security footage of about 1890 hours. It has near about 12 real life anomalies such as vandalism, bomb explosion, thievery, murder etc. All of the above anomalies have much negative impact in the societies and hence needs to be monitored.

2.2 Anomaly Detection Methods

Anomaly detection is considered as detecting some anomalous behavior in the videos. Now there is no proper definition of what an anomaly is, but it can be seen as something which is not usual, normal or predictable. Anomalies should not happen now and then, because it is very difficult for a human to look at so many cameras at the same time and detect anomalies and that too without any error.

The whole idea is to develop a system that automatically detects anomaly behavior. Abnormality can be seen in Figure 1.4 in which the normal or regular regions are identified by N and the abnormal or the unusual behavior is shown by O. As it can be noted from the

figure that all the abnormalities seem to appear outside what is considered normal. Conversely, such aberrations could be close to normality which O2 shows.

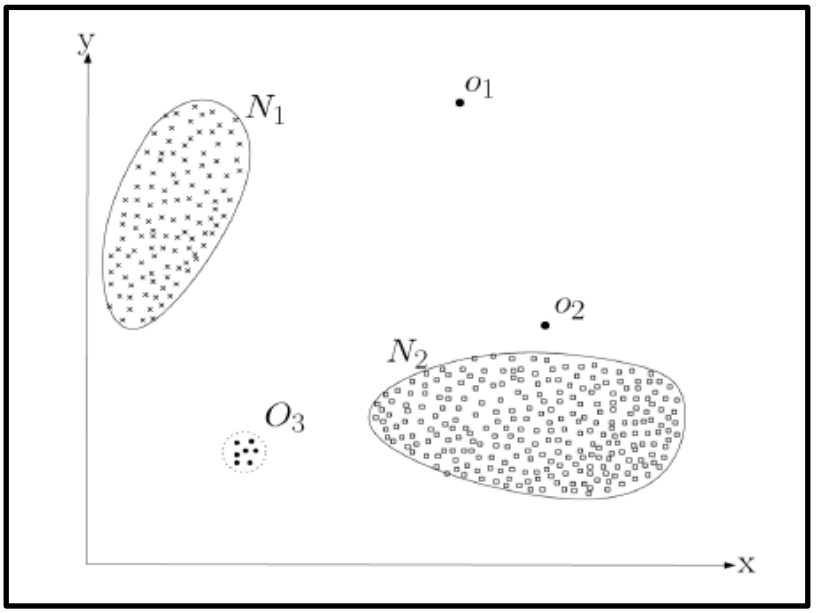


Fig 1.4 A case of anomaly detection

Abnormality detection approaches may be mainly categorized into two areas. Conventional strategies depend simply on clustering-based recognition [11, 12] and low-level retrieval of characteristics.

As deep learning methods have transformed the realm of computer vision [13], all the solutions that incorporate deep learning in their models have shown a tremendous gain in popularity to solve the problem associated with video anomaly detection.

2.2.1 Models Using Trajectories

The whole idea of the methods that use trajectories is based on the fact that all the anomalies are sudden and exhibit abnormal traits across a large variety of video.

Thus, such methods may acquire normal trajectories that occur from normal happenings in a frame. To detect the abnormal behavior, it can be noticed whether the normal behavior or

learned route is obeyed or not.

2.2.2 Models Using Low-Level Feature Extraction

Nevertheless, establishing paths from regular events is hard for standard clustering algorithms. In contrast, to diminish the problem faced in clustering based algorithms that are very much dependent on the motion of the object. The feature extraction of the low-level feature is based on the low-level perception in the video such as fluctuation of greyscale, flowing vectors and textures [14].

2.2.3 Models Using Deep Learning.

There is one most famous method in the deep learning techniques that is solve the video anomaly problems i.e. by using reconstruction error [15, 16]. In this approach, the model is trained for the normal or usual videos such that if there occurs an abnormality, it will always show larger reconstruction errors as compared to the normal frames.

Fundamental frameworks which are popular in image-based models such as Convolution neural networks are used in the video anomaly detection problem.

Now in order to use this model for video analysis instead of pictures or images we need to incorporate temporal features such as LSTM, Two-Stream Model or 3DConvNet.

2.3 Review Of The Recent Work

Detecting abnormalities has always been challenging. There have been several approaches described for recognizing anomalies, which includes detecting anomalies for a given domain such as vehicle accidents [17] and vandalism etc. Such techniques are excellent for the particular reason for which they were intended, but they cannot be extended to other types of domains. In other words, the procedures are not generic and are appropriate in a given context. Anomalies have been identified in a number of methods. In interaction and communication and relationship differences have been studied by the authors, as well as presenting a full description of current technology developments. However, their approach fails to take into account human interaction, and the identification of gestures is also not taken into consideration.

Sultani, et al. [18] suggest learning aberration using the deep multiple instance ranking architecture , where the training labels are at the video level rather than the clip level. They apply MIL, or Multiple Instance Learning, in their technique to dynamically develop a deep anomaly ranking model.

Yu Tian, et al.[19] present Robust Temporal Feature Magnitude Learning (RTFM),which may be deemed quite distinctive as well as a theoretically solid approach for training a feature magnitude learning function. This technique assists in swiftly distinguishing the good examples which in turn enhances the robustness of the MIL strategy against bad instances from aberrant films. This methodology additionally combines diluted convolutions and self attention strategies to capture long- and short-term temporal correlations in order to more precisely understand the feature magnitude.

Somar Boubou and colleagues [20] created a machine learning system for recognising and

interpreting human behaviors. Running, clapping, walking, and side boxing are among the activities. As a consequence, it is evident that the investigation has concentrated entirely on certain human behaviors. However, this strategy is only useful when a single individual is participating in the video. As soon as there are more than one individual, this approach fails to distinguish this tactic. Hyun Seok, Wesley, et al. [21] developed a sparse representation-based technique for human action recognition to characterize the distinctions between the activities. Their technique handles this challenge by recognising irregularities other than human-caused ones, such as automotive wrecks and explosives. J. Kim and K. Grauman [22] established a spatial Markov random field to discover uncommon events. Unusual occurrences are recognised by attempting to compare them to a set of normal models which are established to identify the typical events and if they do not fit, the event is deemed abnormal. This method could perform well if all the activities that are considered normal and non-anomalous is clearly defined and obvious i.e., all routine occurrences could be recognized. We all are aware that there is no appropriate definition of what normal is and so this strategy is not very successful. In reality, "typical" activity sorts are imprecise and not a definite thing. Hence, training a model under normal behavior would be unrealistic as normal is rather a fuzzy word and would be tough to specify all the things that occurs under it.

Sparse coding is another approach that has shown to be beneficial [23,24]. Sparse coding is the encoding of objects by the intense activity of a small number of neurons. This technique builds a vocabulary for normal events from videos that contain normal occurrences. The whole idea of this approach is that whenever an abnormal video will occur, it will create larger reconstruction error, which when compared to the normal videos, will be much higher. Though this strategy is appropriate, surveillance camera video changes over time. As a result, these techniques have a high false alert rate for many routine behaviors.

Paper's Ref	Published on	Supervision	AUC
[25]	ICIP 19	Weakly	78.66
[26]	CVPR 19	Weakly	82.12
[27]	ACM MM 19	Fully	82.0
[28]	ECCV 20	Weakly	83.03
[29]	CVPR 21	Weakly	82.30
[30]	TIP 21	Weakly	84.89

Table 1.1 : Analysis with UCF-Crime Dataset

Above table shows the tabular data of some of the recent papers that have worked on the UCF-Crime Dataset along with their accuracies and the supervision.

Paper's Ref	Published on	Supervision	AUC
[31]	CVPR 21	Unsupervised	90.2
[32]	ICME 2020	Weakly	86.3
[26]	CVPR 19	Weakly	84.44
[28]	ECCV 20	Weakly	89.67
[19]	ICCV 21	Weakly	97.21
[30]	TIP 21	Weakly	97.48

Table 1.2: Analysis with SHANGHAI Tech Dataset

Above table shows the tabular data of some of the recent papers that have worked on the SHANGHAI Tech Dataset along with their accuracies and the supervision.

Paper's Ref	Published on	Supervision	AUC
[33]	ICCV 19	Unsupervised	86.9
[34]	ICME 2020	Weakly	89.2
[35]	ACM MM 20	Unsupervised	90.2
[36]	AAAI 21	Unsupervised	86.6
[37]	TNNLS 21	Unsupervised	88.3
[38]	ACM MM 20	Unsupervised	87.0

Table 1.3: Analysis with AVENUE Tech Dataset

Above table shows the tabular data of some of the recent papers that have worked on the AVENUE Tech Dataset along with their accuracies and the supervision.

CHAPTER 3

PROPOSED WORK

3.1 PROBLEM STATEMENT

With an upswing in crime in today's environment, a smart surveillance system is essential. Delays in responding by competent authorities increase the amount of loss of lives and property. If the right authorities are alerted quickly, the damage could be greatly minimized. Any unusual event in the video stream should be identified by the smart surveillance system in real time without human involvement.

3.2 PROPOSED METHOD

3.2.1 Input Video

CCTV is used to capture what is going on around you. The camera creates a stream of photos that are delivered into the system as input. An individual snapshot may supply us with a plethora of helpful information.

3.2.2 Frame Extraction

In the Frame extraction we have scale each of the frames to 240 * 320 pixels. The video is

made up of numerous frames. For video processing, we employed the FFmpeg programme. We converted the movie into several frames using FFmpeg. The frame rate is set to 30 frames per second.

3.2.3 Feature Extraction

So For video, it is absolutely unfeasible and wasteful to employ the full video as input since it may have numerous similar pictures. Additionally, there are no time explanations for each video. In such circumstances, obtaining essential properties from video and applying them to train the model may be useful.

We implement an I3D feature extractor utilising resnet50 as the base for extracting features. ”Two— Stream Inflated 3D ConvNets” (I3D) is developed on the architecture of computer vision which includes inflating filters and pooling kernels combined. Various computer vision networks such as ResNet, VGG, and InceptionNet are inflated into spatiotemporal feature extractors. These feature extractors create a 2048 feature vector for every video.

3.2.4 Anomalous Behavior Detection

The model for anomaly detection is given the feature vector. After this, It gets determined if the video is anomalous or not. If it's anomalous it will belong to class 0 else class 1.

CHAPTER 4

IMPLEMENTATION

4.1 Dataset Used

We employed the Illegal Surveillance Video dataset given access by UCF. It covers both traditional and odd videos. Ordinary videos are 70GB, whereas odd videos are 20GB. The set comprises 1159 normal films and 950 aberrant ones. We assessed 13 odd instances.

Datasets	Videos in Dataset	Frames in Dataset	Duration
UCSD Ped1	70	201	5 min
UCSD Ped2	28	163	5 min
Subway Entrance	1	121	1.5 hours
Subway Exit	1	64	1.5 hours
Avenue	37	839	30 min
UNM	5	1290	5 min
BOSS	12	4052	27 min
UCF	**1900**	**7247**	**128 hours**

Table 1.4 UCF Crime Dataset

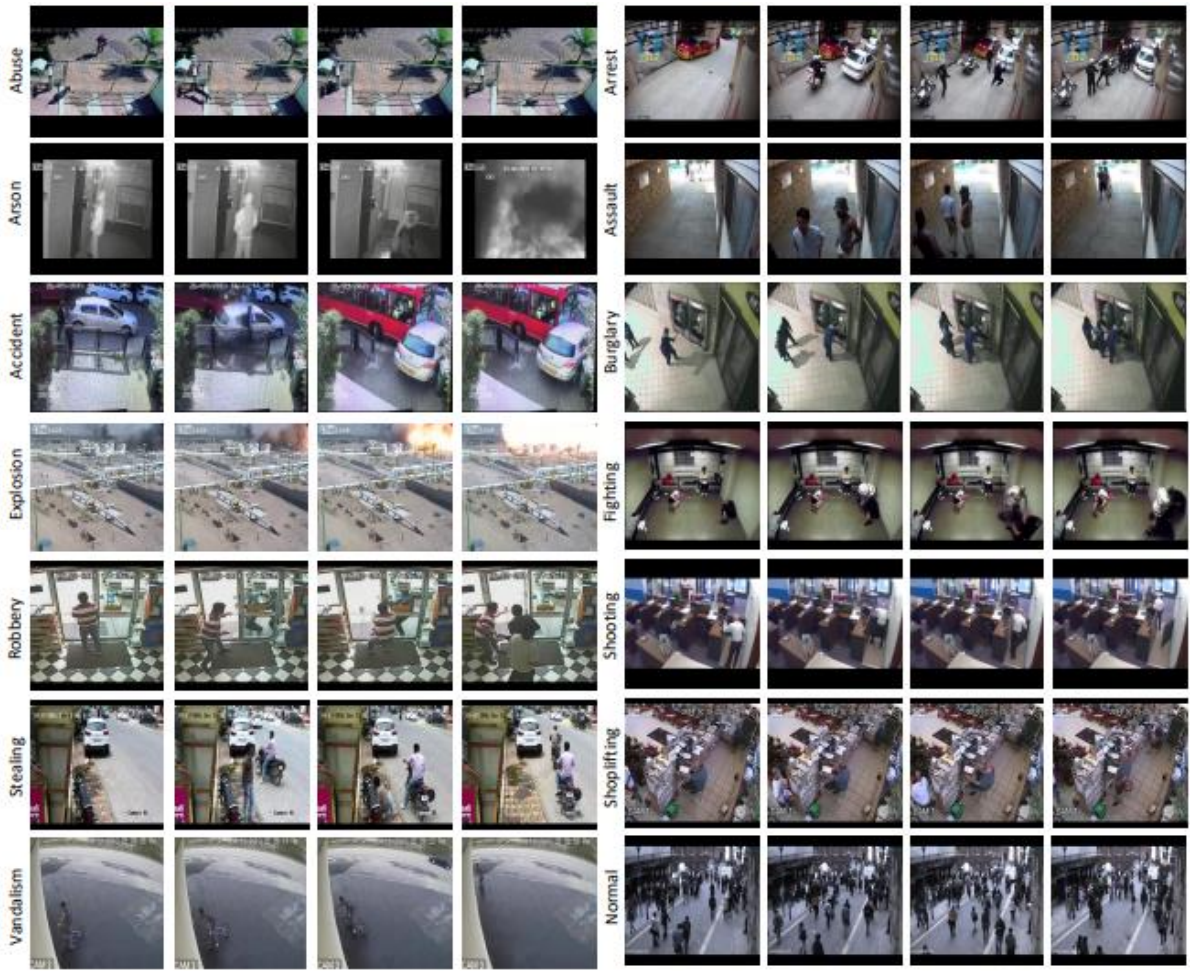


Fig 1.5 Examples of various anomalies in UCF dataset.

Most of the research has explored the amount of frequency of mistakes that is created in model predictions. It is done by employing Equal Error Rate (ERR) measures. It gives the parameters for matching the False Acceptance rate and the False Rejection rate. ERR and model's efficiency are inversely related. The model's accuracy is connected to the Equal Error Rate, If the Equal Error Rate is high then the accuracy of the designed model in considered as low and similarly for the other case. ERR is not only the method to define the accuracy of the model, there is one more method known as F1 Score. F1 Score is also recognized as yet another way for assessing the model's soundness. This technique is used to assess the accuracy of binary categorization. Scholars have developed it to study the reliability of outlier identification, with edge detection improved as 2-class classification. This is obtained by

dividing the amount of recall and accuracy by the sum of the two. It is very different from the ROC curve, it additionally incorporates false positives and false negatives when determining a model's accuracy.

4.2 Implementation

We take a video collection that contains both normal and aberrant movies and feed it into the I3D network. The I3D network then derives the features from the video and puts them in the proper directories. The recovered properties from the I3D network are then utilized to construct the model, which detects aberrant films. The video is submitted into I3D-Resnet50 as an input for feature extraction. The I3D framework is based on resnet50 [39]. By employing a rate which is equivalent to 4 and a cheap 1 x 1 convolution, a bottleneck block is applied to restrict the number of channels. The objective is to lower the number of 3x3 convolution parameters.

Finally, the structure is enhanced with some additional 1x1 convolution. Following each convolutions layer, batch normalization is done, as is the activation function employed is Rectified linear activation function (ReLU). The ResNet50 model is broken into five stages. Each phase consisted of a convolution as well as an Identity block. Each identity block contains three levels of convolution, and the convolution block consists of three convolution layers as well. To train the resnet50, the Kinetics400 dataset is utilized. Weights are transferred to the I3d network by resnet50. The video is later utilized as input by the algorithm.

The footage was pre-processed using RGB and Flow NumPy arrays. RGB and the flow inputs that comprise continual flow of information are utilized to train the I3D network. A two-stream network is useful as optical flow technologies are repeating in certain respects. The

characteristics are retrieved by integrating the outputs from both networks. Following feature extraction, we input the collected features into the CNN architecture.

There are five hidden layers in the model. Following each convolution layer is a dropout [40] layer. To deter the model against overfitting, a dropout rate of 0.5 is utilized. Because of its simplicity and ease of computation, the activation function employed is ReLU. It provides sparsity in the hidden units by giving values ranging from zero to maximum. The final level leverages sigmoid and offers deterministic judgment of the video being abnormal. There are 75 epochs and a training batch size of 64.

4.3 Evaluation Metrics

Various research has been undertaken in this sector and in majority of them it is proposed that AUC and ROC curves are suitable metrics for analyzing the model. Given the prior system's vulnerability to falsification (failure to distinguish rare video) and substantial probability of fraudulent reports, we are additionally utilizing F1-Score as an evaluation measure. We do not utilize Equal Error Rate (EER) as it is useless for recordings that have a small duration of anomalous sections within a longer video.

CHAPTER 5

EXPERIMENTS AND RESULTS

A binary classification model is developed utilizing characteristics collected from I3D resnet. Normal films are labeled as 1, whilst aberrant videos are marked as 0. The algorithm must learn to give more points (nearer to one) to Regular films and significantly fewer scores (nearer to zero) to Anomalous videos. We examined various classification approaches for this task, including the Random Forest Classifier, XG-Boost Classifier, Bagging Classifier, KNN Classifier and CNN.

All of these above mentioned classifier is trained on 1285 films before ever being tested on 506 fresh videos. Figure 1.5 illustrates the test results. All other classifiers were outperformed by the XG-Boost Classifier and CNN. CNN learns more complicated data contained within the features. As a result, it is the most accurate and also has the best f1-score. This seven - layer CNN was trained for 75 epochs utilizing a 33 percent validation split and a batch size of CNN, resulting in a training accuracy of 89 percent and a testing accuracy of 84 percent. We acquired 0.321 drop in training and 0.482 drop in testing utilizing binary cross-entropy as our CNN's loss function.

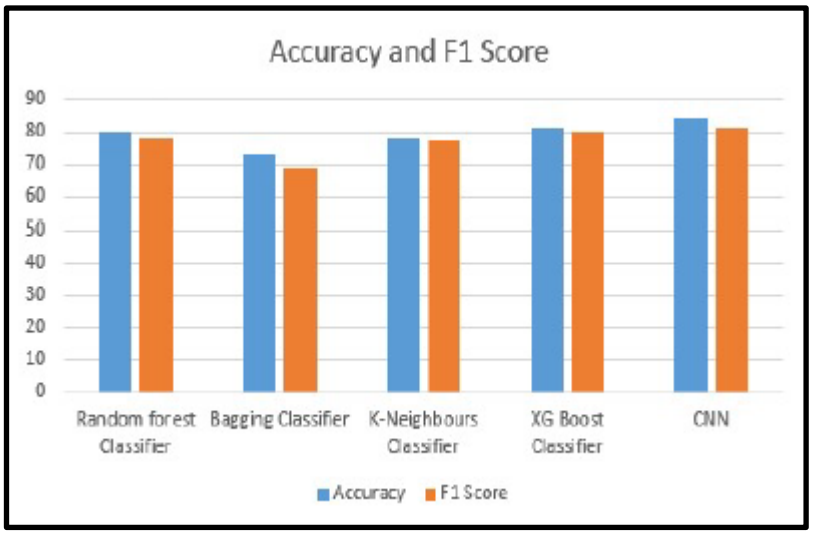


Fig 1.6 F1 Score and Accuracy

Below table gives the tabular information about some of the previous methodologies that have used UCF Dataset along with their accuracies. It clearly shows that our methodologies which uses CNN classifier outperformed all the previous works that have been done using this dataset.

Methodology	**Accuracy**
Fully Auto-encoder [41]	50.3
Sparse Combination Learning [42]	65.49
C3D + Two Stream [18]	75.39
I3D + RGB [19]	84.03
C3D + MIL [18]	75.39
Ours + XG-Boost Classifier	**81.2**
Ours + CNN Classifier	**84.45**

Table 1.5 Performance Comparison of the UCF-Crime Dataset

CHAPTER 6

CONCLUSION

In this study we suggested an algorithm for identifying irregularities in videos used for surveillance. The suggested algorithm has been tested on a variety of surveillance footage and effectively found the abnormality. These results further suggest that feature extraction done by I3D-Resnet-50 removes the demand for temporal annotation. When compared to existing anomaly detection approaches, our methodology performed better. On the UCF-Crime dataset, our system was tested and confirmed, and it reached an accuracy of 84.28 percent.

Our model also generates an F1-Score of 83.05 percent, indicating a low incidence of false alerts in the system. Rather than defining the sort of aberration or moment in the video, our method involves understanding an anomalous event. When this strategy is combined with monitoring technology, it will drastically minimize human error and physical labor.

REFERENCES

[1] A Alfred Raja Melvin, G Jaspher W Kathrine, S Sudhakar Ilango, S Vimal, Seungmin Rho, Neal N Xiong, and Yunyoung Nam. Dynamic malware attack dataset leveraging virtual machine monitor audit data for the detection of intrusions in the cloud. Transac- tions on Emerging Telecommunications Technologies, page e4287, 2021.

[2] Q. Wu, Y. Liu, Q. Li, S. Jin and F. Li, "The application of deep learning in computer vision," 2017 Chinese Automation Congress (CAC), Jinan, 2017, pp. 6522-6527.

[3] Divya Thakur, Rajdeep Kaur.An Optimized CNN based Real World Anomaly Detection in Surveillance Videos.

[4] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. Robust real time unusual event detection using multiple fixed location monitors. PAMI, 30:555 – 560, 2008. ResearchGate, available online: https://www.researchgate.net/publication/330357393

[5] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In CVPR, 2004.

[6] O. Boiman and M. Irani. Detecting irregularities in images and in video. In ICCV, 2005.

[7] N. Jmour, S. Zayen and A. Abdelkrim, "Convolutional neural networks for image classification," 2018 International Conference on Advanced Systems and Electric Technologies (ICASET), Hammamet, 2018.

[8] J. Kim, J. Kim, H. L. T. Thu and H. Kim, "Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection," 2016. International Conference on Platform Technology and Service (PlatCon), Jeju, 2016.

[9] T. Lima, B. Fernandes and P. Barros, "Human action recognition with 3D convolutional neural network," 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), Arequipa, 2017.

[10] S. V. C. Lab, “UCSD anomaly data set”, 2014. “http://www.svcl. ucs d.edu/projects/anomaly/dataset.html”

[11] Jessie James P Suarez and Prospero C Naval Jr. A survey on deep learning techniques for video anomaly detection. arXiv preprint arXiv:2009.14146, 2020.

[12] Muazzam Maqsood, Maryam Bukhari, Zeeshan Ali, Saira Gillani, Irfan Mehmood, Seungmin Rho, and Young Jung. A residual-learning-based multi-scale parallel convolutions- assisted efficient cad system for liver tumor detection. Mathematics, 9(10):1133, 2021.

[13] Yuxuan Zhao, Ka Lok Man, Jeremy Smith, Kamran Siddique, and Sheng-Uei Guan. Improved two-stream model for human action recognition. EURASIP Journal on Image and Video Processing, 2020(1):1-9, 2020.

[14] Siqi Wang, En Zhu, Jianping Yin, and Fatih Porikli. Video anomaly detection and localization by local motion based joint video representation and ocelm. Neurocom- puting, 277:161{175, 2018.

[15] Davide Abati, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Latent space autoregression for novelty detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 481{490, 2019.

[16] Joey Tianyi Zhou, Jiawei Du, Hongyuan Zhu, Xi Peng, Yong Liu, and Rick Siow Mong Goh. Anomalynet: An anomaly detection network for video surveillance. IEEE Trans- actions on Information Forensics and Security, 14(10):2537{2550, 2019.

[17] S. Mohammadi, A. Perina, H. Kiani, and M. Vittorio. Angry crowds: Detecting violent events in videos. In ECCV, 2016.

[18] Sultani, Waqas, Chen Chen, and Mubarak Shah. ”Real-world anomaly detection in surveillance videos.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[19] Y. Tian, G. Pang, Y. Chen, R. Singh, J. W. Verjans and G. Carneiro, ”Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning,” 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 4955-4966, doi: 10.1109/ICCV48922.2021.00493.

[20] Somar Boubou and Einoshin Suzuki. 2015. Classifying actions based on histogram of oriented velocity vectors. J. Intell. Inf. Syst. 44, 1 (February 2015), 49–65. https://doi.org/10.1007/s10844-014-0329-0

[21] Y.-L. Hou and G. K. Pang, “People counting and human detection in a challenging situation,” IEEE transactions on systems, man, and cybernetics-part a: systems and humans, vol. 41, no. 1, pp. 24–33, 2011.

[22] J. Kim and K. Grauman. Observe locally, infer globally: A space-time mrf for detecting abnormal activities with incremental updates. In CVPR, 2009.

[23] B. Zhao, L. Fei-Fei, and E. P. Xing. Online detection of unusual events in videos via dynamic sparse coding. In CVPR,2011.

[24] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In ICCV, 2013.

[25] J. Zhang, L. Qing and J. Miao, ”Temporal Convolutional Network with Complementary Inner Bag Loss for Weakly Supervised Anomaly Detection,” 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 4030-4034, doi: 10.1109/ICIP.2019.8803657.

[26] J. Zhong, N. Li, W. Kong, S. Liu, T. H. Li and G. Li, ”Graph Convolutional Label Noise Cleaner: Train a Plug-And-Play Action Classifier for Anomaly Detection,” 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1237-1246, doi: 10.1109/CVPR.2019.00133.

[27] Kun Liu and Huadong Ma. 2019. Exploring Background-bias for Anomaly Detection in Surveillance Videos. In Proceedings of the 27th ACM International Conference on Multimedia (MM ’19). Association for Computing Machinery, New York, NY, USA, 1490–1499. DOI:https://doi.org/10.1145/3343031.3350998

[28] Zaheer, Zaigham Mahmood, Arif Astrid, Marcella Lee, Seung-Ik. (2020). CLAWS: Clustering Assisted Weakly Supervised Learning with Normalcy Suppression for Anomalous Event Detection.

[29] J. -C. Feng, F. -T. Hong and W. -S. Zheng, ”MIST: Multiple Instance Self-Training Framework for Video Anomaly Detection,” 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14004-14013, doi: 10.1109/CVPR46437.2021.01379.

[30] P. Wu and J. Liu, ”Learning Causal Temporal Relation and Feature Discrimination for Anomaly Detection,” in IEEE Transactions on Image Processing, vol. 30, pp. 3513-3527, 2021, doi: 10.1109/TIP.2021.3062192.

[31] M. -I. Georgescu, A. B˘arb˘al˘au, R. T. Ionescu, F. Shahbaz Khan, M. Popescu and M. Shah, ”Anomaly Detection in Video via Self-Supervised and Multi-Task Learning,” 2021 IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), 2021, pp. 12737-12747, doi: 10.1109/CVPR46437.2021.01255.

[32] B. Wan, Y. Fang, X. Xia and J. Mei, ”Weakly Supervised Video Anomaly Detection via Center-Guided Discriminative Learning,” 2020 IEEE International Conference on Multimedia and Expo (ICME), 2020, pp. 1-6, doi: 10.1109/ICME46284.2020.9102722.

[33] T. N. Nguyen and J. Meunier, ”Anomaly Detection in Video Sequence With Appearance-Motion Correspondence,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1273-1283, doi: 10.1109/ICCV.2019.00136.

[34] Liu, W., Luo, W., Li, Z., Zhao, P., Gao, S. (2019). Margin Learning Embedded Prediction for Video Anomaly Detection with A Few Anomalies. IJCAI.

[35] Yu, Guang Wang, Siqi Cai, Zhiping Zhu, En Xu, Chuanfu Yin, Jianping Kloft, Marius. (2020). Cloze Test Helps: Effective Video Anomaly Detection via Learning to Complete Video Events. 10.1145/3394171.3413973.

[36] Ruichu Cai, Hao Zhang, Wen Liu, Shenghua Gao, Zhifeng Hao: Appearance-Motion Memory Consistency Network for Video Anomaly Detection. AAAI 2021: 938-946

[37] Wang, X., Che, Z., Yang, K., Jiang, B., Tang, J., Ye, J., Wang, J., Qi, Q. (2021). Robust Unsupervised Video Anomaly Detection by Multi-Path Frame Prediction. IEEE transactions on neural networks and learning systems, PP.

[38] Ziming Wang, Yuexian Zou, and Zeming Zhang. 2020. Cluster Attention Contrast for Video Anomaly Detection. In Proceedings of the 28th ACM International Conference on Multimedia. Association for Computing Machinery, New York, NY, USA, 2463–2471.

[39] K. He, X. Zhang, S. Ren and J. Sun, ”Deep Residual Learning for Image Recognition,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 770-778.

[40] Srivastava, Nitish Hinton, Geoffrey Krizhevsky, Alex Sutskever, Ilya Salakhutdinov, Ruslan. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research. 15. 1929-1958.

[41] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury,and L. S. Davis. Learning temporal regularity in video sequences. In CVPR, June 2016.

[42] C. Lu, J. Shi, and J. Jia. Abnormal event detection at 150 fps in matlab. In ICCV, 2013.

LIST OF PUBLICATIONS

[1] R. K. Yadav and R. Kumar, "A Survey on Video Anomaly Detection," **2022 IEEE Delhi Section Conference (DELCON)**, 2022, pp. 1-5, doi: 10.1109/DELCON54057.2022.9753580.

Indexed by **Scopus** and **Google Scholar**

Published in IEEEXplore

URL:

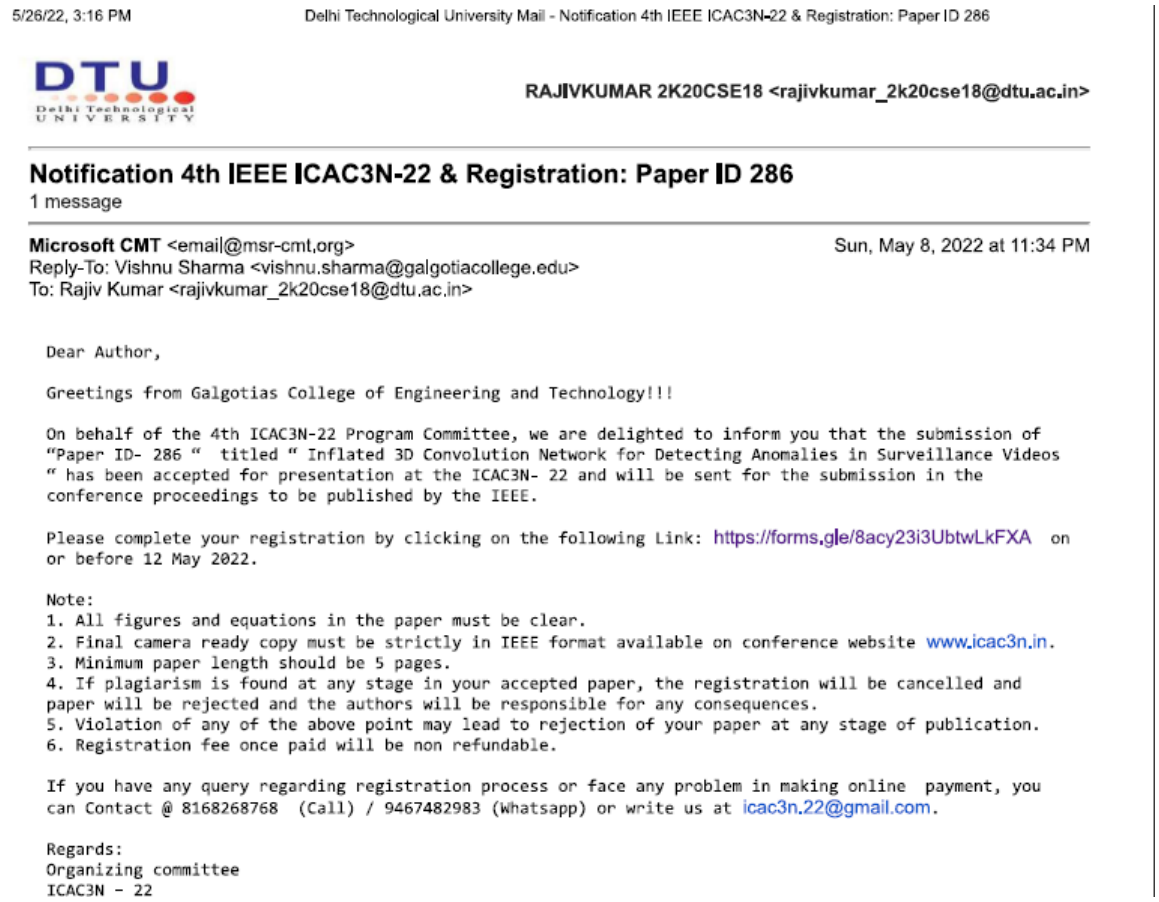
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9753580&isnumber=9752688

CERTIFICATE

[2] R.K. Yadav and Rajiv Kumar, “Inflated 3D Convolution Network for Detecting Anomalies in Surveillance Videos”. **Accepted** at the **4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N–22).**

Indexed by **Scopus** and **Google Scholar**

PROOF OF ACCEPTANCE



5/26/22, 3:16 PM Delhi Technological University Mail - Notification 4th IEEE ICAC3N-22 & Registration: Paper ID 286

DTU Delhi Technological University

RAJIVKUMAR 2K20CSE18 <rajivkumar_2k20cse18@dtu.ac.in>

Notification 4th IEEE ICAC3N-22 & Registration: Paper ID 286

1 message

Microsoft CMT <email@msr-cmt.org> Sun, May 8, 2022 at 11:34 PM
Reply-To: Vishnu Sharma <vishnu.sharma@galgotiacollege.edu>
To: Rajiv Kumar <rajivkumar_2k20cse18@dtu.ac.in>

Dear Author,

Greetings from Galgotias College of Engineering and Technology!!!

On behalf of the 4th ICAC3N-22 Program Committee, we are delighted to inform you that the submission of “Paper ID- 286 “ titled “ Inflated 3D Convolution Network for Detecting Anomalies in Surveillance Videos “ has been accepted for presentation at the ICAC3N- 22 and will be sent for the submission in the conference proceedings to be published by the IEEE.

Please complete your registration by clicking on the following Link: https://forms.gle/8acy23i3UbtwLkFXA on or before 12 May 2022.

Note:
1. All figures and equations in the paper must be clear.
2. Final camera ready copy must be strictly in IEEE format available on conference website www.icac3n.in.
3. Minimum paper length should be 5 pages.
4. If plagiarism is found at any stage in your accepted paper, the registration will be cancelled and paper will be rejected and the authors will be responsible for any consequences.
5. Violation of any of the above point may lead to rejection of your paper at any stage of publication.
6. Registration fee once paid will be non refundable.

If you have any query regarding registration process or face any problem in making online payment, you can Contact @ 8168268768 (Call) / 9467482983 (Whatsapp) or write us at icac3n.22@gmail.com.

Regards:
Organizing committee
ICAC3N – 22