# A Spatial-Temporal based Next Frame Prediction and Unsupervised Classification of Video Anomalies in Real Time Estimation

**Swapna Kumari Sahu**, **M. Jayanthi Rao**

***Abstract: Anomaly detection is an area of video analysis has a great importance in automated surveillance. Although it has been extensively studied, there has been little work started using CNN networks. Hence, in this thesis we presented a novel approach for learning motion features and modeling normal Spatio-temporal dynamics for anomaly detection. In our technique, we capture variations in scale of the patterns of motion in an image object by using optical flow dense estimation technique and train our auto encoder model using convolution long short term memories (ConvLSTM2D) as we are processing video frames and we predict the anomaly in real time using Euclidean distance between the generated and the ground truth frame and we achieved a real time accuracy of nearly 98% for the youtube videos which are not used for either testing or training. Error between the network's output and the target output is used to classify a video volume as normal or abnormal. In addition to the use of reconstruction error, we also use prediction error for anomaly detection. The prediction models show comparable performance with state of the art methods. In comparison with the proposed method, performance is improved in one dataset. Moreover, running time is significantly faster.***

***Keywords: Spatio-Temporal video features, ConvLSTM2D***

## I. INTRODUCTION

As video analytics play a major role in security analytics and most of the countries fund widely for the security applications using Artificial Intelligence (AI), many researchers bring forth their products which works well with the train and test datasets but fails in real time environments due to camera angles, bad weather occlusions in cameras or due to hefty noisy environments and uneven indoor/outdoor light conditions. Although the recent closed circuit capture camera fills up with technologies like pixel enhancement, auto correction in focus, resizing, orienting, and color corrections and despite of many preprocessing techniques has been come to the practice such as sift, surf, etc., and histogram based feature extraction, image filtering and image derivative techniques such as Laplacian, Gaussian and Laplacian over Gaussian (LoG) and famous deep learning models such as Resnet50, InceptionNet and VGG16/19 used as a image feature extraction technique, the models were built with less real time accuracy even with the test dataset. Actually, the anomaly detection is a part of video classification which in turn the internal architecture separates into three approaches such as Motion based input features, 2Dmodel and 3D model as said by Caleb Andrew [1] et al., in his review paper on video classification techniques. He also stated that,Thomas G. Dietterich et al., came up with the Multiple Instance Technique (MIS) [2][3][4][5][6] for the training samples which are not clear (hence weakly supervised) and he also stated that Annabella Astorino et al., proposed an architecture for the image classification based on MIS model which emphasizes the current models of binary classification technique and also he explained that the current machine learning model combines both MIS and Conditional Random Fields (CRF) methods for video classification in which the whole video is taken as a bag and the individual frames of each videos are considered as the instances. And most of the latest research papers uses thishybrid technique which is a graph based neural networks. And instead of frames as instances theweakly supervised neural network trains feature vectors with the help of graph theory and decision boundary. And these feature vectors are obtained from the previously trained video classification models such as either **"i3d"** or **"c3d"** features. This kind ofMulti Instance Learning (MIL) model for anomaly video classification is proposed by Waqas Sultani [4] in her research paper "Real-world Anomaly Detection in Surveillance Videos", in which she aggregated the temporal segments into two bags such as positive and negative bags and applied "c3d" feature extraction for the temporal features and trained it with the multi-layered network in order to get the instance scores of the temporal segments in the positive and negative bags. Later on, the Multiple Instance Model is utilized by Boysng Wan[1] et al in which i3d features as utilized as instances in the positive and negative bags, and anomaly regression has been done on those bags based on CNN where the dynamic multi instance learning loss (DMILL) is used as a metric for prediction of anomaly in the video frames. Jia-Chang Feng [2] et al., also proposed a weighted classification network which takes the input from the multi instance learning graph that generates the pseudo labels from the graph.

Manuscript received on September 22, 2021.
Revised Manuscript received on September 29, 2021.
Manuscript published on October 30, 2021.
*Correspondence Author

**Swapna Kumari Sahu*,** PG Research Scholars, Department of Computer Science and Engineering Sri Sivani College of Chilakapalem, (Andhra Pradesh) India. Email. swapnasahubtechssit@gmail.com

**Dr. M. Jayanthi Rao,** M. Tech, Ph. D Associate Professor, Department of Computer Science and Engineering, Sri Sivani College of Chilakapalem, (Andhra Pradesh) India. Email. jayanth.mtech@gmail.com

*Retrieval Number: 100.1/ijeat.A31611011121*
*DOI: 10.35940/ijeat.A3161.1011121*
*Journal Website: www.ijeat.org*

*Published By:*
*Blue Eyes Intelligence Engineering*
*and Sciences Publication (BEIESP)*
*© Copyright: All rights reserved.*

A deep temporal encoding-decoding mechanism for the multi instance learning bags is done by Ammar Mansoon Kamoona [3] et al., where he did the same thing like c3d feature extraction and keeping the instances in positive and negative bags and used a temporal classification neural network using convolution neural networks with sigmoid as a predictive function. Jia-Xing Zhong[5] et al., also proposed a Graph Based Convolution Network in which the snippet features are extracted from the action classifier and fed them to two graph modules, in which the first one is the Feature Similarity graph and the second one is the Temporal consistency graph whose outputs are fused and fed to the sigmoid function for binary classification. A holistic neural network was proposed by Peng Wu [6] in which anomaly detection is not only done with the help of video frames but also with the help of audio features which is merged with the video features and given as the input to the graph based holistic network. In his research, he compared his dataset statistics with the other datasets but the main concern with his dataset is, it cannot be used at a place where clumsy background with huge noise is possible in the video segments. And in real time the camera height may be like 9 to 12 feet on road, where the events like accident detection can be recorded but the voice cannot be heard. So the aggregation dataset of voice and video cannot provide maximum target accuracy at various conditions and camera angles and at various instances where occlusion might be not good. A binary entropy based temporal classification model is proposed by YuTian [7] et al. where he proposed two architectures with different feature extraction techniques (“C3D” and “I3D”) where these temporal features are classified based on their magnitudes. This similar kind of feature extraction technique (I3D) for every frame is done and the batches of these features have been taken into account using Sliding window technique [8] across each frame and a Convolution Neural Network is trained on the batches using a combination of both “mean square error (mse)” and “custom loss (anomaly detection loss)” for prediction of loss. A Combination of Person Detection and U-Net Classification with mse as metric is done by Kemal Doshi [9] et al. In his research paper, he is trying to detect the person,whenever an anomaly is done using Yolo object detection and the U-net classification is done by comparing the original frame with the U-net generated image with the help of a threshold. Memory guided normality for anomaly detection with the help of U-net architecture is proposed by Hyunjong Park [10] et al. where the reconstruction of the video is done by mapping frames into the memory in the U-Net architecture and finally the classification is done with the help of L2 distance. Instead of Memory Mapped mechanism a similar kind of architecture with auto encoders and U-Net architecture were proposed by Tong-Nguyen [11][12] in his both research papers for anomaly detection in video sequence. In the paper [12], he used U-Net for instant motion detection in a frame. A Conventional image processing technique called background subtraction from an image using Gaussian Mixture Model and later feature extraction were performed from the fusion of convolution neural network and lstm networks and the classification of these features is done through SVM classifier for prediction of anomaly is also done [13] in which accuracy is somewhat low as the training videos are quite lessto perform with a better test accuracy. Hence, we proposed a model that takes magnitude based Spatio temporal characteristics from the images as input and a 3D Neural Network architecture with time series based convolutions and Convolution LSTM2D architecture and achieved an overall accuracy of more than 99% in the test data and nearly 98% in the validation data which is taken from the youtube videos. The Training dataset is taken from UCF Crime dataset in which each category of anomaly and normal videos or merged and each video is trimmed to 3 secs so as to get 7 frames equally from each video, and the anomaly part of the video contains both anomaly and the end of the normal event, so our network will know whatever the next frame will be coming. In this paper, the below section II will explain the dataset preparation and the architecture of our model,whereas the results, comparison will be continued in the following Section III which in turn followed by Conclusion and Future Scope, Bibliography.

## II. DATASET PREPARATION AND MODEL DESIGN

In the preparation of the dataset we have merged all the UCF Crime dataset and trimmed each and every video to 3 secs so as to get a rate of 7 frames for each video. For the videos which have the anomaly, we kept the starting frames normal and then merged the anomaly part so as to get the next frame prediction in our unsupervised model. In the process, we collected a sample of nearly 100 thousand video samples of both normal and anomaly dataset and trained them with our model up to 20k epochs to get the good accuracy of nearly 98% for real time videos of different anomalies at different camera angles.

The below functional flow describes each and individual step of our algorithm.

### Algorithm

**Input:** merging all the videos of UCF dataset

**Step1.** Clipping each video with a time gap of 3 seconds so as to obtain the total no. of frames of 7 for each video

$$V = (\sum_{1}^{7} v) \left\langle \begin{matrix} Total\ Video\ Samples \\ k = 1 \end{matrix} \right\rangle$$

Here the total video samples are 100 thousand samples from all the classes.

**Output:**

**Step2.**A magnitude based function operator is defined for each video frame such that the magnitude operator is negative if the LoG of the current frame magnitude is less than some threshold value otherwise it is kept positive.

$$f(x) = \begin{cases} -v, & x < 0\ (label\ 0) \\ v, & x \geq 0\ (label\ 1) \end{cases}$$

Here ‘x’ is the direction of the magnitudes in each frame (v) which can be found using optical flow estimation technique and if the LoG of all the magnitudes of a frame is negative then we label the frame as ‘0’ else ‘1’

**Step3.**After labeling each and every frame based on the their magnitudes for every video we then mark the final label as **“anomaly”** if the number of “1$^{s}$” are more otherwise we label it as **“normal”**and based on the number of positive frames we later decided whether the anomaly comes under **low risk** or **high risk** factor.

**Step4.** Training labeled dataset with the model up to 20k epochs with **binary cross entropy (bce)** as loss function and **Adam**as optimizer so the model trains faster with a better accuracy.

**Step5.**While prediction, we loop over for every 7 frames and predict each frame with our model so whenever the Euclidian distance between the generated frame and the original frame differs more than 50% then we come to the conclusion as the frame has anomaly.

## III. FUNCTIONAL BLOCK DIAGRAM

The below figure depicts the function block diagram of our complete architecture. In the figure, we process 7 frames from every 3 seconds video and obtain the magnitudes for every frame using dense optical flow mechanism and apply LoG function so as to get the label of the frame from the magnitude. Then we trained our data by using time series architecture for frame regeneration so as to predict the next frame of the video using Euclidean distance.

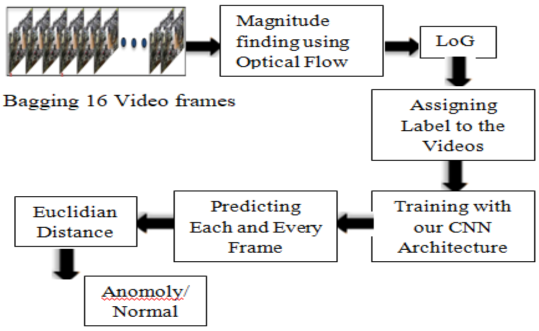

Magnitude finding using Optical Flow
LoG
Bagging 16 Video frames
Assigning Label to the Videos
Training with our CNN Architecture
Predicting Each and Every Frame
Euclidian Distance
Anomoly/ Normal

**Fig1. Block Diagram of our model**

## IV. PROPOSED ARCHITECTURE

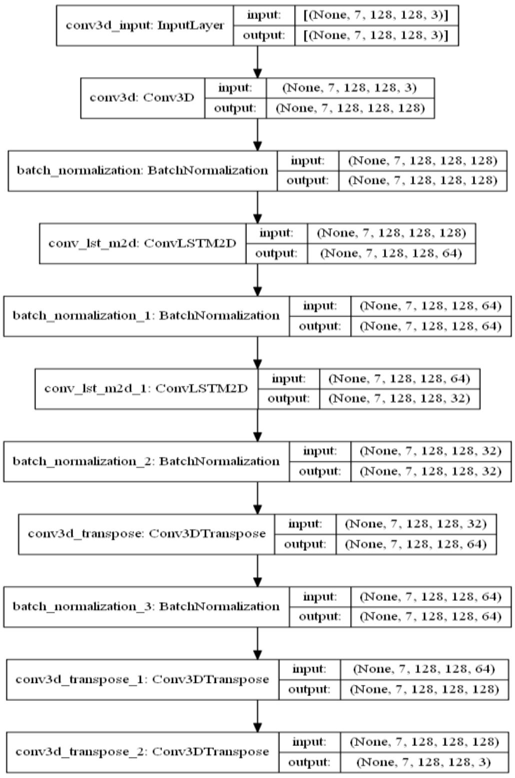

conv3d_input: InputLayer | input: | [(None, 7, 128, 128, 3)]
output: | [(None, 7, 128, 128, 3)]
conv3d: Conv3D | input: | (None, 7, 128, 128, 3)
output: | (None, 7, 128, 128, 128)
batch_normalization: BatchNormalization | input: | (None, 7, 128, 128, 128)
output: | (None, 7, 128, 128, 128)
conv_lst_m2d: ConvLSTM2D | input: | (None, 7, 128, 128, 128)
output: | (None, 7, 128, 128, 64)
batch_normalization_1: BatchNormalization | input: | (None, 7, 128, 128, 64)
output: | (None, 7, 128, 128, 64)
conv_lst_m2d_1: ConvLSTM2D | input: | (None, 7, 128, 128, 64)
output: | (None, 7, 128, 128, 32)
batch_normalization_2: BatchNormalization | input: | (None, 7, 128, 128, 32)
output: | (None, 7, 128, 128, 32)
conv3d_transpose: Conv3DTranspose | input: | (None, 7, 128, 128, 32)
output: | (None, 7, 128, 128, 64)
batch_normalization_3: BatchNormalization | input: | (None, 7, 128, 128, 64)
output: | (None, 7, 128, 128, 64)
conv3d_transpose_1: Conv3DTranspose | input: | (None, 7, 128, 128, 64)
output: | (None, 7, 128, 128, 128)
conv3d_transpose_2: Conv3DTranspose | input: | (None, 7, 128, 128, 128)
output: | (None, 7, 128, 128, 3)

**Fig2.Architecture of our proposed Network**

The above picture depicts our autoencoder model architecture, where the input data was given to a 3d convolutional neural network and later the output of the unit is fed to the batch normalization layer to provide some regulation and to reduce the internal covraition shift at the end of every epoch. And at every layer we used a non-linear kernel initializer called **“he-norm”**so as to start the weights from the standard Gaussian distribution and we used rectification linear unit (relu) as the activation function for all the layers except the last layer as our data is non-linear and Gaussian distributed. The input shape of our first 3d convolution layer is (7, 128, 128, and 3) such that every video consists of only 7 frames with a shape of W*H (128, 128) and depth of 3 which means we are taking a color frame as an input. As our model requires prediction of anomaly from a time distribution sequence, I have used a Convolution Long short Term memory module for processing of the data which is a sequence of images (Video). Later we used the 3d transpose of the convolution layer so as to retrieve the sequence of frames after being processed from the previous stages. In the 3d convolution, the filter moves across the (x, y, z) directions and generates a cuboid of 128 samples in the 3d space as per our architecture. The stride I have used is (1, 1, 1) such that every pixel detail will be processed in all the three directions so as to ensure low less as our architecture is very small due to the account of big dataset taken into consideration. The below graphs depicts the train, test accuracy and losses for all the 20k epochs

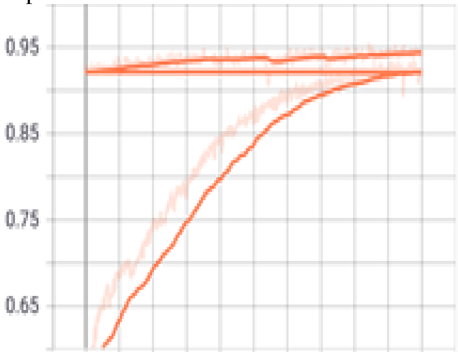


**Fig3. Training Accuracy (for 20k epochs)**

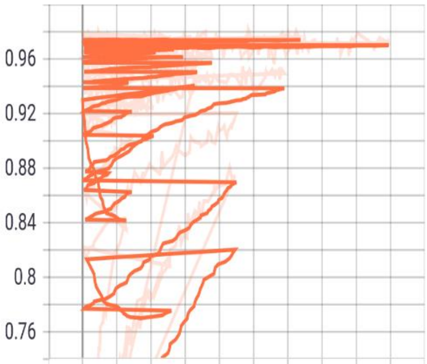


**Fig4. Testing Accuracy (for 20k epochs)**

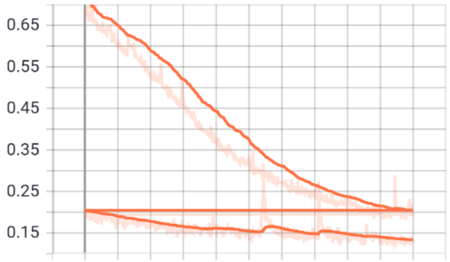


Fig5. Training Loss (for 20k epochs)

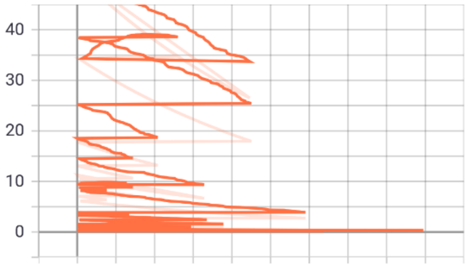


Fig6. TestingLoss (for 20k epochs)

As each anomaly videos is a mixture of normal as well as anomaly frames and the classification is unsupervised (label names are giving by taking LoG of the magnitudes taken from the optical estimation) and the network architecture is low, some gradients during training and testing stuck at local minima so most part of the graph has deviation even the Adam optimizer is used with a proper learning rate of 0.001. Hence we increased the number of epochs to 20k so as to achieve a good accuracy of nearly 98% for the train data and 95% for the test data. But even though we got a test accuracy of 95% the validation accuracy was nearly 97% for all the collected videos from the youtube which are the real time scenarios at different camera angles. Some of the image output has been displayed below from the youtube videos

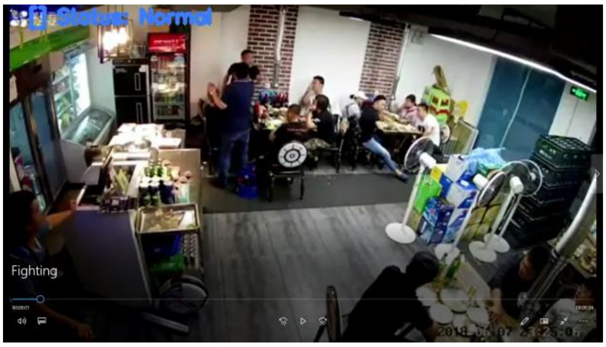


Fig7. Persons sitting in a Restaurant (Normal Frame)

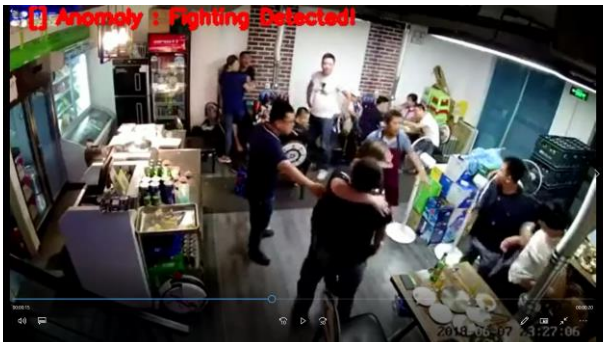


Fig8. Anomaly Detected when person fighting in a video (Labeled: Fighting in the video for clarification)

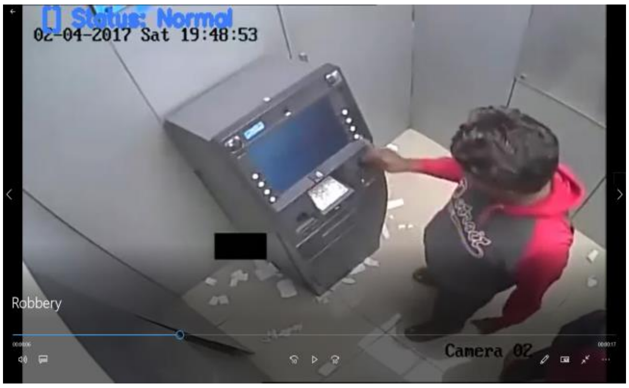


Fig 9. Person Drawing Cash from the ATM

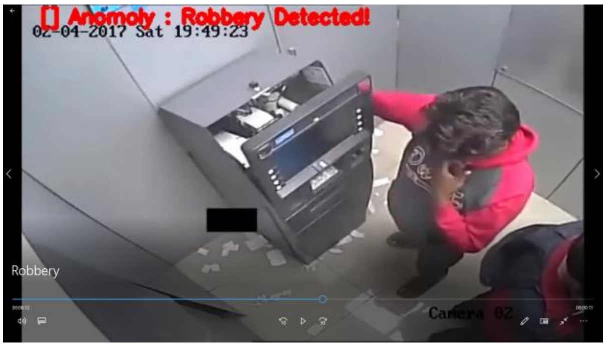


Fig10. Anomaly Detected (Labeled: ATM Tampering in the video for clarification)

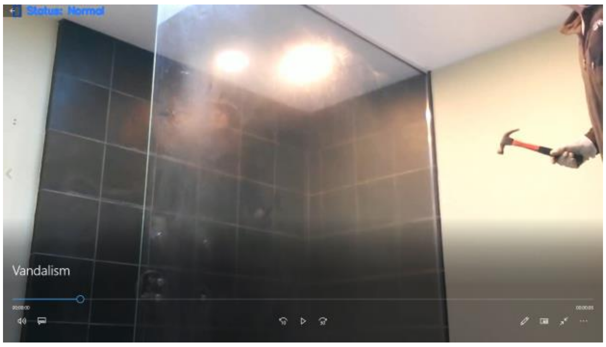


Fig11. Normal Video

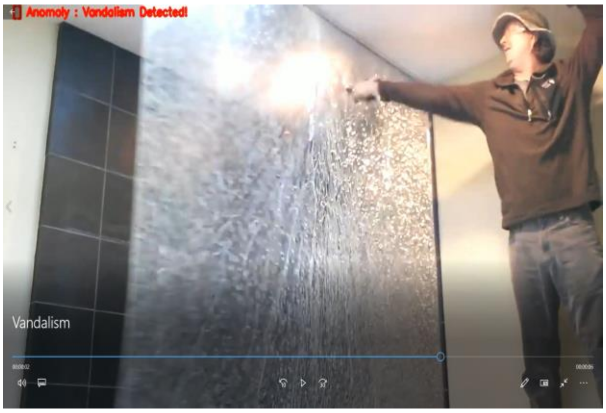


Fig12. Person Destroying the Glass in his washroom with a small hammer (Labeled: Vandalism Detected in the video for clarification)

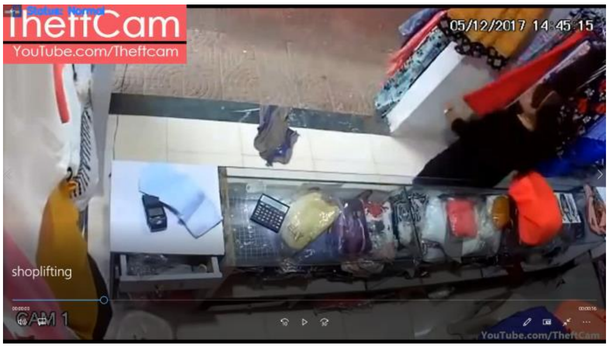


**Fig13. Person waiting in a clothes shop**

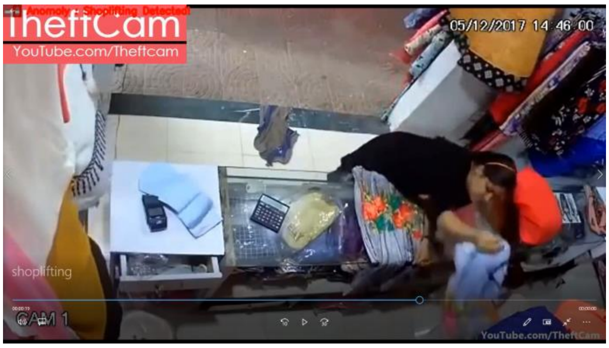


**Fig13. Person Stealing clothes from the clothes shop (Labeled: Shoplifting Detected in the video for clarification)**

The above frames are collected from youtube videos and tested with our live model with different camera angles and different anomalies, hence we labeled anomaly name at the test side for the clarification purpose.

## V. CONCLUSION ANF FUTURESCOPE

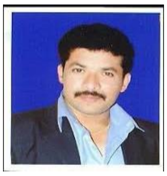
Our magnitude based online labeling of UCF Crime dataset along with pure unsupervised modeling of anomaly detection is performed quite well with the unseen youtube videos taken at different angles. And we achieved a maximum performance of 98% of training, 95% of Testing and 97% of the Validation Data and we want to enhance our network architecture by astrous convolution layers instead of 3d convolution layer at the beginning and at the end of our architecture and want to verify its performance.

## REFERENCE

1. A survey on video classification using action recognition, Caleb Andrew et al. International Journal of Engineering & Technology
2. Weakly supervised video anomaly detection via center-guideddiscriminative learning, BOYANG WAN et al. ©2020 IEEE
3. MIST: Multiple Instance Self-Training frameworkforVideo Anomaly Detection, Jia-Chang FengIEEE International Conference on Computer Vision and Pattern Recognition. 2021.
4. Multiple Instance-Based Video Anomaly Detectionusing Deep Temporal Encoding-Decoding, Ammar Mansoor Kamoona et al. arXiv: 2007.01548v2
5. Real-world Anomaly Detection in Surveillance Videos, Waqas Sultani et al. arXiv:1801.04264v3
6. Graph Convolutional Label Noise Cleaner: Train a Plug-and-play Action Classifier for Anomaly Detection, Jia-Xing Zhong et al, arXiv:1903.07256v1
7. Not only Look, but also Listen: Learning Multimodal Violence Detection under Weak Supervision, Peng Wu, Jing Liu et al. arXiv: 2007.04687v2
8. Weakly-supervised Video Anomaly Detection with Robust Temporal Feature Magnitude Learning, Yu Tianwt al. arXiv: 2101.10030v3
9. ADNet: Temporal Anomaly Detection in Surveillance Videos, Halil Ibrahim Ozturk et al, arXiv: 2104.06653v1
10. Online Anomaly Detection in Surveillance Videos withAsymptotic Bounds on False Alarm Rate, Keval Doshi et al, arXiv:2010.07110v1
11. Learning Memory-guided Normality for Anomaly Detection, Hyunjong Park et al, arXiv: 2003.13228v1
12. Hybrid Deep Network for Anomaly Detection, Trong-Nguyen Nguyen et al, arXiv: 1908.06347v1
13. Anomaly Detection in Video Sequence with Appearance-Motion Correspondence, Trong-Nguyen Nguyen et al, arXiv:1908.06351v1
14. Abnormal Event Detection on BMTT-PETS 2017 Surveillance Challenge, Kothapalli Vignesh et al. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops
15. X. Cui, Q. et al. Abnormal detection using interaction energy potentials. In Computer Visionand Pattern Recognition (CVPR), IEEE, 2011

## AUTHORS PROFILE

**Miss. Swapna Kumari Sahu,** is a post graduate student in the field of Computer Science Engineering. She has an interest in the field of image & video analytics done as a research in her Post Graduate Program



**Dr.M.Jayanthi Rao**, has done his Ph.D in Nagarjuna University, currently acting as Head of the Department in the field of Computer Science Engineering has a vast knowledge in Image and Video Analytics